Innovation and Change in Professional Education 13

Paul F. Wimmers
Marcia Mentkowski *Editors*

Assessing Competence in Professional Performance across Disciplines and Professions

With a Foreword by Lee S. Shulman

Springer

Innovation and Change in Professional Education

Volume 13

SCOPE OF THE SERIES

The primary aim of this book series is to provide a platform for exchanging experiences and knowledge about educational innovation and change in professional education and post-secondary education (engineering, law, medicine, management, health sciences, etc.). The series provides an opportunity to publish reviews, issues of general significance to theory development and research in professional education, and critical analysis of professional practice to the enhancement of educational innovation in the professions.

The series promotes publications that deal with pedagogical issues that arise in the context of innovation and change of professional education. It publishes work from leading practitioners in the field, and cutting edge researchers. Each volume is dedicated to a specific theme in professional education, providing a convenient resource of publications dedicated to further development of professional education.

More information about this series at http://www.springer.com/series/6087

Paul F. Wimmers · Marcia Mentkowski
Editors

Assessing Competence in Professional Performance across Disciplines and Professions

With a Foreword by Lee S. Shulman

Springer

Editors

Paul F. Wimmers
University of California
Los Angeles, CA
USA

Marcia Mentkowski
Alverno College
Milwaukee, WI
USA

Innovation and Change in Professional Education
ISBN 978-3-319-30062-7   ISBN 978-3-319-30064-1 (eBook)
DOI 10.1007/978-3-319-30064-1

Library of Congress Control Number: 2016932870

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Foreword

Forewords are a delightfully open-ended and forgiving genre of composition. They can be discursive and analytic, offering a conceptual introduction to the volume that follows them. They can also serve as an excuse for personal observations, in which a writer who did not contribute to the volume itself offers observations and anecdotes that might bring readers of the book to approach its contents at a more personal level. I have chosen to indulge in a bit of both options. My conceit is that in the intersection of personal experiences and conceptual/technical challenges lays a narrative that is useful in understanding where we stand and how we got there. I hope to be forgiven for the effort.

For the past 60 years (!), I have nurtured an ongoing and occasionally contentious relationship with the varieties of assessment. The relationship began when, at the age of 16, I spent a week responding to a daunting set of placement examinations, primarily in multiple choice and essay formats, administered to all undergraduates admitted to the College of the University of Chicago. Those placement exams were samples of the final exams of the yearlong general courses of the College's curriculum. Several weeks after the tests administration, I learned that I had passed the equivalent of the first two years of the social science curriculum and two-thirds of the first year of the humanities course. I needed only to study first-year art appreciation and art history, and I would be eligible to begin the second year of humanities.

A few years later, as a doctoral candidate, I was already feeling dissatisfied with the limitations of multiple-choice tests however well designed (and the Chicago tests were outstanding examples of the genre). I spent a year as research assistant to one of the university's examiners, Christine McGuire, who was moonlighting at the Center for the Study of Liberal Education for Adults. We created assessments of critical thinking and judgment for adult students in Chicago's Great Books extension program. The assessments asked them to "perform" critical thinking as if they were members of juries hearing cases, or public officials making judgments about competing policies, after having read Plato's *Republic* or Mill's *On Liberty.* The assessments asked respondents to make sequential judgments; identifying information they needed, gathering that information by rubbing out locations on data sheets, then making new judgments as they moved to their judgments and decisions. These assessments were forms of performance assessment that would later be elaborated by McGuire in the development of "Patient Management Problems" in the assessment of medical expertise. The very next year, I became research assistant to a professor who had served as one of McGuire's mentors in the university examiner's office, Benjamin Bloom.

I became enamored of these ways of creating task environments that permitted us to tap into thought processes (I had, after all, written a bachelor's thesis on approaches to epistemology.). I designed a doctoral dissertation in which I studied how teachers solved classroom problems by creating a simulation of an entire classroom captured in an "in-basket" assessment. The assessment focused on how the teachers discovered and formulated problems rather than how they solved problems already framed. To study problem finding required the use of more unstructured task situations. Although Bloom was my academic advisor for several years, I was more attracted to cognition than to measurement, and thus never became a psychometrician.

Another few years passed into the late 1960s and early 1970s, and as a young professor of educational psychology and medical education at Michigan State University, my colleagues and I were designing and using "high-fidelity" simulations to study and assess the ability of medical students to make complex diagnoses. Michigan State's brand-new medical school had opted to open its doors in 1968 with a "focal problems" curriculum that used clinical problems rather than disciplines or organ systems as the basis for the curriculum. In our studies we used a variety of assessment instruments and situations—from written patient management problems to simulations engaging actors and actresses to represent clinical cases. Each encounter was videotaped and then used in stimulated recall to probe deeper into the knowledge, problem-solving skills and even the interpersonal empathy and doctor–patient relations of the aspiring physicians. Many of these high-fidelity methods became part of the Emergency Medicine specialty assessments that we designed, and have now found their way into most medical education programs in America as both instruction and assessment methods.

In the mid-1980s, now at Stanford, my students and I responded to the challenge of developing much more valid approaches to the assessment of high levels of pedagogical competence among experienced teachers. We first developed multi-day assessment centers that looked quite similar to what would, in medical education, become OSCEs. Although our assessments met all of the psychometric standards we established (with Lee J. Cronbach and Edward Haertel as our hard-nosed measurement collaborators), we moved beyond those methods and developed yearlong situated portfolio assessments so that the effects of real settings and changes over time could become part of the assessment picture. These studies led to the development of the *National Board of Professional Teaching Standards* certification exams that remain in use today.

And in the first decade of the present century, while at the Carnegie Foundation, we studied the processes of professional education in a variety of fields—law, engineering, medicine, nursing, the clergy, and business. In all of these, issues of professional assessment loomed large, whether we examined the uses of student evaluations over the course of training or at the manner in which assessment for licensure was conducted.

The Cognitive… and Beyond

One of the most important things we learned in our Carnegie studies of learning in the professions was that we needed to think of the development of professional learning as "learning to profess." Professing entailed more than deep understanding. This kind of learning comprised three distinctive yet interacting kinds of learning. Learning to profess involved the development of habits of mind, habits of practice, and habits of the heart. All three taken together were the necessary features of professional learning. This conception of "learning to profess" places significant burdens on the developers of assessments.

Most of the traditional advances in assessment had dwelled on deepening and enriching the ways in which intellectual prowess could be measured. My teacher Ben Bloom had emphasized, in his taxonomies, the importance of moving beyond knowledge to applications, from analysis to synthesis and critical evaluation, from comprehension to problem solving, from reproducing knowledge to creating understanding. The big challenges were to elicit those varieties of understanding, and to do so in objective, reliable, reproducible, and demonstrably fair ways because the results of those exams were used consequentially for making admissions, graduation, and licensure and certification decisions.

The assessment of habits of practice, with the intellectual and technical *skills* of the professions, did not work so well with multiple-choice or essay exams, so performance assessments had to be created. The experiences I had with Christine McGuire prefigured the current generations of creative assessments—from the CLA in liberal arts to computer-based simulations in medicine, engineering and the military, and to direct observations of students working on complex tasks in context.

But what of habits of the heart? What of the moral, ethical and formational attributes associated with professional integrity and identity? As we examined professional learning across diverse fields, we concluded that the development of "professional identity" was paramount in those domains. Mind, hand and heart are integrated in professional identity formation that is critical in guiding the manner in which professionals make judgments and decisions under the conditions of uncertainty that characterize their work. These kinds of assessment remain among the most difficult of challenges, and perhaps ultimately they can only be observed carefully within mentored practice settings rather than "measured" by some test-like event. It is, indeed, this kind of challenge that has led to Alverno's commitment to its highly elaborated system of portfolio- and performance-based assessments embedded in their programs. It also was the reason that our work on the National Board assessments for experienced teachers led us to portfolio-based assessments in that field.

Alverno and Maastricht: A Lovely Juxtaposition

I find it particularly fitting that the co-editors of this volume, Marcia Mentkowski and Paul Wimmers, are associated with two institutions whose characters are so intimately associated with the insight that assessment must be integrated with curriculum and instructional program if it is to become a powerful influence on the educational process.

Marcia Mentkowski has been a leader of the remarkable pedagogical accomplishments of Alverno College since joining its faculty in 1976. Marcia immediately began to integrate evaluation and research with the evolving innovations of that small Catholic women's college that has emerged as one of the world's outstanding laboratories for ambitious teaching, learning, and assessment. I remember the excitement I felt when I first learned about the work at Alverno. That excitement has not abated. The college's leaders wrote about their work, presented at national meetings, and conducted empirical studies of its efforts. Mentkowski's work at Alverno exemplifies the value of a scholarship of teaching and learning that strengthens one's own institution simultaneously providing both inspiration and empirically based insights to an entire field.

At about the same time that Mentkowski arrived at Alverno, Maastricht University opened its doors in the Netherlands for the first time in January of 1976. I had learned about the plans for Maastricht seven years earlier when I served for a month as a visiting professor in the University of Leiden School of Medicine. I was there because in my faculty role at the Michigan State University College of Human Medicine, we had opened our doors with a "focal problems" curriculum, a problem-based approach to the teaching of medicine. At Leiden, I met a young physician, Evert Reerink, who was soon to become one of the pioneers who founded the Maastricht program that so deftly integrated problem-based learning and the kinds of assessment needed to sustain such instruction. The co-editor of this volume, Paul Wimmers, worked, studied, and conducted research at Maastricht before moving to UCLA. The spirit of Maastricht permeates his research and writing. And like Alverno, Maastricht has become a great source of the literature on problem-based curricula and the kinds of performance assessment that must accompany them.

Why did Alverno College in Milwaukee and the Maastricht School of Medicine in the Netherlands become the unlikely sites of such important innovations? The histories of these two institutions teach us about the importance of context in the world of reform in higher education. Alverno is an urban Catholic liberal arts college for women in Milwaukee, Wisconsin. When the charismatic Sister Joel Read became president in 1968, she brought a vision of academic excellence and vocational relevance rooted in the humane Catholic values of education as formation. It was consistent with the institution's history and commitments. Alverno accepted most of the young women who applied, rather than selectively winnowing the applicants before they had stepped inside. In a nonselective school committed to the success of its students, the quality and character of the teaching was critically important, and assessment must be used to guide and support success rather than to sort students out once they had been admitted. The Alverno program of small classes that were highly practice-oriented, low stakes, and highly informative assessment embedded and threaded throughout the curriculum, was essential and it was accomplished. Alverno College thus became the site for the most mature forms of assessment in the service of teaching and learning, an institution that could exploit the potential of assessment to exemplify what is worth learning—meeting the ultimate challenge of measuring educational formation.

Maastricht has a parallel history 4000 miles away. In a country that already had seven fine medical schools, the decision to build an eighth in Limburg was contentious. The response was not to build a clone of Leiden or Amsterdam, but to design something utterly new, experimental, and audacious. As a new school, they attracted exciting, visionary new faculty members who had been inspired by new visions of medical education in places like McMaster and Michigan State that also had flexibility because they did not have to cope with the burdens of high prestige. So they invented a university, not only a school of medicine, using problem-based learning and having to invent new forms of assessment that would be faithful to the novel curriculum. Here again, as at Alverno, the context invited innovation.

Because of Alverno's commitment to practical relevance for its young women, and Maastricht's commitment to clinical practice, the most important standards of assessment design were often not the classical psychometric principles of reliability and validity, but the often denigrated principles of fidelity to the real world, face validity, and congruence with the flow of the curriculum. Both institutions were at least a generation ahead of their time. They have been joined by a number of others, though the inertia of higher education remains difficult to overcome.

University of Chicago and the Uses of Assessment

Permit me a personal reminiscence. When I was an undergraduate in the College of the University of Chicago in the latter half of the 1950s, none of the students in the College realized we were part of a remarkable experiment in academic assessment. Indeed, evaluation was considered so important to the academic life of the College, that a totally independent office of evaluation had been established under the aegis of President Robert Maynard Hutchings to design and manage the evaluation of undergraduates. The three successive university examiners read like a history of psychological and educational measurement: L.L. Thurstone, Ralph Tyler, and Benjamin Bloom.

Using assessment diagnostically invites the danger of excusing students from courses from which they could profit greatly, even though they demonstrate the requisite knowledge and skills that the course is designed to impart. In my case, I pursued a career in the social sciences, but missed the chance to study with David Reisman, who taught in the second year social science core course. Would studying with Reisman have changed the kind of social scientist I would become? We will never know.

On the other hand, the placement exam results placed me in a four-student class in art history and appreciation, which left an indelible mark on my love of the visual arts. The class even supported a three-week liaison with El Greco's *Assumption of the Virgin* in the Art Institute, an encounter that taught me so much about esthetic design that I later taught a class about the renal system to first-year medical students using the *Assumption* as my analogical inspiration.

There were several features of student assessment at Chicago that are relevant to the topics of this volume. First, student achievement was understood as a set of performances that began with "knowledge" but did not end there. Under the leadership of university examiner Benjamin Bloom and his colleagues, a "taxonomy of educational objectives" was elaborated that distinguished lower order thinking—knowledge, comprehension, and application—from higher order thinking—analysis, synthesis, and evaluation. Since these were understood as distinctive, albeit interdependent, performances, different kinds of test items and item sequences were designed to test them. All courses were expected to provide learning opportunities at all these levels, assessment blueprints were designed to map those courses accordingly, and assessments were designed to correspond to those maps.

Second, a clear distinction was made between assessment for the purposes of guiding learning and providing students with information on their development, and assessment for assessing students' achievements for the purpose of assigning final grades "for the record." The system was quite extreme. Undergraduate courses lasted for an entire academic year. They were peppered with tests and written assignments to provide formative feedback to students, but none of those performances were ever made part of the student's academic records. They were entirely formative. A student's final grade for each yearlong course was based on a 6–9 h comprehensive examination—both multiple-choice and essay—undertaken at the end of May. A student could retake that comprehensive examination in any field later in his or her undergraduate years, and the last performance would be the basis for the grade on their academic record.

Indeed, so seriously were the performances on these assessments taken, that placement examinations were administered to all entering freshmen before their first year's courses. On the basis of those placement exams, students who did quite well were given course credit and relieved of the obligation to take those courses. Since the undergraduate curriculum at Chicago was defined as 14 yearlong courses, some students were able to complete their bachelor's degrees in far less than 4 years.

There is a critical distinction to be made between high- and low-stakes assessments. Too often, in the current climates of assessment in the service of accountability, external assessments carry quite high stakes and are rarely of much diagnostic or formative value to teachers or learners. I refer to this genre of assessment as “high stakes/low yield” and alas, they have come to dominate the world of assessment (though typically not in the work of the authors of this volume). The ideal use of assessment strives to create “low stakes/high yield” evaluations that provide useful, just-in-time information to those who teach and learn, and do not carry so much weight “for the record” that they invite teaching-to-the-test, the gaming of assessments, and the corruption of education by assessment.

When assessment is low stakes/high yield, it is embedded into the flow of instruction, and it is difficult, even unnecessary, to distinguish between teaching and testing. Even though assessment may be inaccurate, the two activities serve the same purpose of guiding, informing, and enriching the experience of learning. Ultimately, some form of summative assessment may be necessary for earning credit, graduation, certification, or licensure. Those of us responsible for assessment, however, must be mindful and responsible in understanding what we are doing and when.

At Chicago, the low stakes/high yield comprehensive examination system led to the use of “no stakes/high yield” student evaluation during the 8 months of active reading and seminars, followed by examinations that were not designed for the faint of heart.

In the most sophisticated of current approaches to the assessment at Alverno and Maastricht, the worst feature of the Chicago model is avoided. That is the feature of an external assessment that has the couple of weeks of “high stakes/high anxiety” comprehensive virtue of not interfering with the diagnostic and instructional value of embedded assessments, but avoids the intense pressure and associated anxiety associated with a single high-stakes test that can be emotionally stressful and even crippling. I believe that Alverno’s combination of Franciscan caring and its commitment to the emotional and social development of its student body at least as much as to their intellectual development has led to this tough but humane hybrid form of assessment and instruction.

Summary Observations

This book reads like the product of a different culture, a different civilization, and a different worldview from the volumes on testing and measurement that characterized the psychometrics 50 years ago. For this we can be grateful. There is much more practical judgment involved, more reliance on skilled multiple reviews than on one-shot high-stakes assessments. As Lee Cronbach advised me when we worked together on the National Teaching Board assessments, these practical scholars do not permit the tail of psychometrics to wag the dog of sound assessment and quality teaching.

Many authors of the chapters in this volume wrestle impressively with some of the most vexing challenges of education. They reject what many educators view as distinctions and contradictions. They design assessments for both professions *and* disciplines, measuring competence *and* performance, balancing assessment *and* instruction, and they recognize that all those distinctions stand alternately as synonyms and antonyms, as analogues and contrasts, as tensions and as coordinate phrases.

I believe that the deep message of this stimulating set of papers is both integrative and analytic. As I read the book (and of course, I read through lenses that have been ground to the prescription of my own idiosyncratic perspectives as they developed over the past 50 years), professions and disciplines are more similar than they are alike, even though their signatures indeed must be distinguished one from the other. Competence and performance must be closely connected or those who profess will be in danger of behaving like one-handed piano players, and even together the two capacities are dangerously incomplete without the integration of integrity and the synthesis of identity. And unless all forms of assessment are ultimately pedagogical and all forms of instruction are laced with assessment for the benefits of both those who teach and those who learn, teaching and learning will be unsafe, unwise, and irresponsible.

Stanford, CA Lee S. Shulman
July 2015

Contents

Editors and Contributors

About the Editors

Dr. Paul F. Wimmers Ph.D. received his M.S. degree at Maastricht University, the Netherlands, in cognitive and educational psychology in 2002. He completed his Ph.D. research in expertise in medicine and the development of clinical competence at the Psychology Department of Erasmus University Rotterdam in 2005. In 2005 he accepted a faculty position in the Office of the Vice Dean at the University of Pittsburg School of Medicine where he continued to work on research in medical expertise. In 2006 he relocated to UCLA, School of Medicine, to accept his current position. He is now Professor of Medicine and Director of Evaluations at the David Geffen School of Medicine at UCLA.

In order to support scholarship and educational leadership among students and faculty, he is Course Chair of the Medical Education Research Selective, Chair of the fourth year teaching fellowship, and Co-chair of the two-year Medical Education Fellowship (MEF) program.

In his research, Dr. Wimmers favors the application of multivariate analysis and factor analysis using techniques like structural equation modeling in the study of expertise, problem solving, problem-based learning, and professional learning and assessment.

Dr. Marcia Mentkowski Ph.D. is a Senior Scholar for Educational Research, Professor Emerita Psychology, and Founding Director Emerita of the Educational Research and Evaluation Department at Alverno College.

Prior to these distinguished roles, Dr. Mentkowski served as Associate Professor at the University of Toledo, Invited Visiting Scholar at Harvard University, and Lecturer at the University of Wisconsin, Milwaukee.

Dr. Mentkowski has consulted on curriculum in law, dentistry, medicine, management, engineering, and judicial education, and brings conceptual direction and administrative expertise to the development and study of student learning outcomes through creating a collaborative culture of evidence from educational theory, research, assessment, and educational policy. She has participated in

12 consortia made up of nearly 170 institutions, and in a wide range of grants and projects funded by private and public entities. Dr. Mentkowski received the prestigious Kuhmerker Award in 1985, and the highest award granted by Lawrence University—the Lucia Russell Briggs Distinguished Achievement Award—in 2012.

Dr. Mentkowski received her MA and Ph.D. in Educational Psychology from the University of Wisconsin at Madison, and BA in Educational Psychology from Downer College of Lawrence University.

Contributors

Erika Abner L.L.B. (York) L.L.M. (York) Ph.D. (Ontario Institute for Studies in Education/University of Toronto) was appointed to the position of Faculty Lead, Ethics and Professionalism for undergraduate medical education at the University of Toronto in January 2014. She has taught and designed curriculum in law school courses, the Ontario Bar Admission Course, and continuing legal education. She is also engaged in research in the practice of law and has published two articles with Shelley Kierstead of Osgoode Hall Law School on legal writing. In 2013, they completed a research paper on Learning Professionalism in Practice, funded by the Chief Justice of Ontario's Advisory Committee on Professionalism. Her research interests include workplace learning, learning in the early years of professional practice, and the differences between novice and expert legal writers. She teaches curriculum development as an Adjunct Lecturer with the Dalla Lana School of Public Health.

Jeana Abromeit Ph.D. is a Professor of Sociology, Social Science Department, Alverno College. She is also Associate Vice President for Academic Affairs. Her research interests include curriculum and pedagogy, program and institutional assessment, oral history research, gender, cultural competence, social stratification, and social conflict. Drawing on her expertise in these areas, she has provided consultation services to a variety of organizations over the past 35 years. She has extensive experience as a consultant for higher education institutions in the U.S. and internationally on teaching, learning, and assessment. For further contact, see jeana.abromeit@alverno.edu.

Susan Baillie Ph.D. is an Adjunct Associate Professor, and Director of Graduate Medical Education at the David Geffen School of Medicine at UCLA. She is responsible for overseeing all 71 graduate medical education programs at UCLA and reports directly to the Senior Associate Dean of Student Affairs. Dr. Baillie has worked extensively with faculty development programs for on-campus and community-based faculty and with the integration of new curriculum components. She is the educational representative to the UCLA National Center for Excellence in Women's Health, and is co-editor of the Women's Primary Care Guide, a 300-page pocket guide on women's health for clinicians. She also writes on medical education and graduate medical education.

Jodi Cohn Dr.P.H. is the Director, Research and Planning, HealthCare Services at SCAN Health Plan in Long Beach, CA. Her area of responsibility is to research evidenced-based interventions in caring for older people, and apply these practices to the SCAN care management program. Dr. Cohn's other area of emphasis is creating practical tools for SCAN's contracted medical providers. She has many years of experience in planning and evaluating health and community-based services for older people. Dr. Cohn received her doctoral degree in Public Health from the UCLA School of Public Health, Division of Health Services.

Ronald Cohn Ph.D. is a Senior Consultant with the Ralston Consulting Group, an organization development firm whose primary focus is helping organizations manage the change process. For over 30 years, he has specialized in team building, work process redesign, executive coaching, leadership development, third-party conflict interventions, small- and family-run business transitions, strategic planning, executive selection, mergers and acquisitions, and helping all levels of employees deal with the messiness of change. Prior to joining Ralston Consulting, Dr. Cohn served as Dean, College of Education, Westminster College, and a full Professor of Education, Psychology, and Business. Dr. Cohn received his Bachelor's degree in Philosophy and History from Clark University and his doctorate in Clinical and Developmental Psychology from the University of Northern Colorado.

Mary E. Diez Ph.D. is Professor Emerita at Alverno College. She served as Dean of the School of Education and Graduate Dean, among other roles at the College. A recognized leader in teacher education nationally and internationally, she has published research and policy works in teaching, learning, and assessment, teaching standards, and teacher development. She currently serves as President of the School Sisters of St. Francis, a congregation of women religious in 11 countries.

Steven J. Durning M.D., Ph.D. is a Professor of Medicine and Pathology at the Uniformed Services University (USU). He received his M.D. degree from the University of Pittsburgh and he practices general internal medicine. He received his Ph.D. from Maastricht University, which addressed the influence of contextual factors on clinical reasoning. Dr. Durning currently oversees a second-year medical students' course on clinical reasoning. In addition to serving as a course director, he is the Director of the newly established Masters and Ph.D. in Health Professions Education at USU and is the Principal Investigator of USU's Long Term Career Outcome Study. Dr. Durning has published over 200 peer-reviewed manuscripts, 20 book chapters, and five books. Dr. Durning serves on a number of national and international organizations and his research interests include clinical reasoning and assessment.

Lily Fountain M.S., R.N., C.N.M. is an Assistant Professor of Nursing at the University of Maryland School of Nursing in Baltimore, MD. Her research focuses on critical thinking in nursing, active learning strategies, and maternal newborn health outcomes research. For publications and contact information: http://www.nursing.umaryland.edu/directory/lily-fountain/.

Cha-Chi Fung Ph.D. is Vice-Chair of the Department of Medical Education and Assistant Dean of Educational Affairs at Keck School of Medicine of USC. Her area of expertise lies in the teaching and assessment of clinical performance and clinical reasoning. Dr. Fung is Chair-Elect for the AAMC Western Group on Educational Affairs and a facilitator and member on the Steering Committee of the Medical Education Research Certificate program sponsored by the AAMC.

Larry D. Gruppen Ph.D. is a Professor in the Department of Learning Health Sciences at the University of Michigan Medical School, where he directs the competency-based Master in Health Professions Education program. His research interests center around the development of expertise, knowledge and performance assessment, self-regulated learning, and educational leadership development. He has over 120 peer-reviewed publications on a variety of topics in medical education and presents regularly at national and international professional meetings. He was recognized for his career productivity by the AAMC's Central Group for Educational Affairs' Medical Education Laureate Award and the 2015 John N. Hubbard Award from the National Board of Medical Examiners. For publications and contact information, see http://lhs.medicine.umich.edu/people/larry-d-gruppen.

Lourdes R. Guerrero Ed.D., M.S.W. is an Adjunct Assistant Professor of Medicine at the David Geffen School of Medicine at UCLA. She has worked in Graduate Medical Education and the Clinical and Translational Science Institute, focusing on research education, training, and career development programs (CTSI-ED). Lourdes completed her doctorate in Educational Leadership at UCLA and holds degrees from UC Berkeley and the Catholic University of America in Washington, DC. Her research agenda includes health disparities and public policy, evaluation of educational programs, diversity and workforce development, and women in academic health sciences.

Ilene Harris Ph.D. is a Professor, Head and Director of Graduate Studies in the Department of Medical Education at the University of Illinois College of Medicine-Chicago. Her research interests focus on assessment in clinical settings, curriculum studies, and qualitative methods. Her research has been reported in over 135 publications in peer-reviewed journals and over 300 presentations at national and regional meetings. She has had major leadership roles in health professions education. For example, she was Chair of the Association of American Medical Colleges (AAMC) national Research in Medical Education (RIME) conference and Chair of the AAMC Central Region Group on Educational Affairs (CGEA). In the American Educational Research Association (AERA) Division of Education for the Professions (Division I), she was President and prior to that Secretary and Program Chair. In recognition of her scholarship and leadership in professions education, in

2010 she received the AERA Division I Distinguished Career Award, an award given by the division to one individual every 2 years.

John Heywood Ph.D. is a Professorial Emeritus of Trinity College Dublin, the University of Dublin. His primary interest is in education for the professions, especially engineering and teacher education. He was awarded the best research publication award of the division for the Professions of the American Educational Research Association in 2006 for his book "*Engineering Education: Research and Development in Curriculum and Instruction*" published by Wiley/IEEE. He is co-author of *Analysing Jobs* (Gower Press)—a study of engineers at work. His other publications include *Assessment in Higher Education, Student Learning, Programmes and Institutions* (Jessica Kingsley), and *Learning Adaptability and Change*: *The Challenge for Education and Industry* (Paull Chapman/Sage). He is a Fellow of the American Society for Engineering Education and a Life Senior Member of the Institute of Electrical and Electronic Engineers.

Brian D. Hodges M.Ed., Ph.D., M.D., F.R.C.P.C. is a Professor in the Faculty of Medicine and the Faculty of Education (OISE/UT) at the University of Toronto, the Richard and Elizabeth Currie Chair in Health Professions Education Research at the Wilson Centre for Research in Education and Vice President Education at the University Health Network (Toronto General, Toronto Western, Princess Margaret, and Toronto Rehab Hospitals). He leads the AMS Pheonix Project: A Call to Caring, an initiative to rebalance the technical and compassionate dimensions of healthcare.

Eric Holmboe M.D. a board-certified internist, is Senior Vice President, Milestones Development and Evaluation at the Accreditation Council for Graduate Medical Education (ACGME). From 2009 until January, 2014 he served as the Chief Medical Officer and Senior Vice President of the American Board of Internal Medicine and the ABIM Foundation. He originally joined the ABIM as Vice President for Evaluation Research in 2004. He is also Professor Adjunct of Medicine at Yale University, and Adjunct Professor of Medicine at the Uniformed Services University of the Health Sciences and Fineberg School of Medicine at Northwestern University.

Prior to joining the ABIM in 2004, he was the Associate Program Director, Yale Primary Care Internal Medicine Residency Program, Director of Student Clinical Assessment, Yale School of Medicine and Assistant Director of the Yale Robert Wood Johnson Clinical Scholars program. Before joining Yale in 2000, he served as Division Chief of General Internal Medicine at the National Naval Medical Center. Dr. Holmboe retired from the US Naval Reserves in 2005.

His research interests include interventions to improve quality of care and methods in the evaluation of clinical competence. His professional memberships include the American College of Physicians, where he is a Master of the College, Society of General Internal Medicine and Association of Medical Education in Europe. He is an honorary Fellow of the Royal College of Physicians in London.

Dr. Holmboe is a graduate of Franklin and Marshall College and the University of Rochester School of Medicine. He completed his residency and chief residency at Yale-New Haven Hospital, and was a Robert Wood Johnson Clinical Scholar at Yale University.

Lois Kailhofer is a professor of mathematics at Alverno College. She is currently serving as chair for the Mathematics and Computing Department and the chair of the Problem Solving Ability. She is active on the Assessment Council at Alverno college.

Shelley Kierstead Ph.D. has been the Director of Osgoode Hall Law School's Legal Process course since 2002. She also teaches family law and child protection law at Osgoode Hall Law School. Her research interests lie in the areas of professionalism, the development of expertise in legal writing, children and families, and therapeutic jurisprudence. She and Erika Abner have published two articles in the legal writing area, as well as a research report on Learning Professionalism in Practice. The research for that report was funded by the Chief Justice of Ontario's Advisory Committee on Professionalism. Professor Kierstead is also a co-author of two texts on legal research, writing, and analysis.

Catherine Knuteson Ph.D., R.N. is a Professor of Nursing at Alverno College. As a School of Nursing faculty member, she teaches both undergraduate and graduate courses, makes clinical placement arrangements, and chairs the college's Social Interaction Ability committee.

Ming Lee Ph.D. an Associate Professor of Medicine at David Geffen School of Medicine at UCLA, is a trained educational psychologist with extended years of experience in medical education and research. Her research interests include clinical performance assessment, development and validation of assessment instruments, program evaluation, and humanistic medicine. She has been serving on a number of journal article review boards in the fields of medical education and research, including *Academic Medicine*, *Medical Education*, and *Journal of American Geriatrics Society*. Her engagement in teaching spans undergraduate and faculty development levels, covering broad content areas such as end-of-life care, precepting challenging students, interprofessional team care, educational assessment tools, objective structured clinical examination (OSCE), and how to construct NBME-style test items.

Dena Lieberman Ph.D. is a Professor of Business and member of the Problem Solving Department and Council for Student Assessment at Alverno College, Milwaukee, Wisconsin, USA. She is experienced in the development of learning and assessment materials for undergraduate business and management curriculum and has served as a consultant to educational institutions on assessment design.

Desiree Pointer Mace Ph.D. is an Associate Professor and Associate Dean for Graduate Programs in the School of Education at Alverno College, and the author of "Teacher Practice Online: Sharing Wisdom, Opening Doors" (Teachers College

Press 2009; http://tinyurl.com/276u9fb). Dr. Pointer Mace's work focuses on envisioning and inventing ways of representing teaching and learning using new media and online technologies, and advocating for high educational outcomes for all students. Dr. Pointer Mace is the conceptual architect and multimedia curator of Inside Mathematics (http://www.insidemathematics.org), a project of the Noyce Foundation. Dr. Pointer Mace received her B.A. in Cognitive Science from Vassar College and an M.A. and Ph.D. in Education with a concentration in language, literacy, and culture from UC Berkeley.

Suzanne Mente M.S. is Assistant Director of Instructional Services at Alverno College where she teaches mathematics. Her interests center around strengthening undergraduate students' quantitative literacy ability through work with students, faculty, and the National Numeracy Network. For contact information, see www.alverno.edu

Heather Mernitz Ph.D. is an Associate Professor of Physical Science and Chair of the Council for Student Assessment at Alverno College, Milwaukee, WI, USA. Her scholarly interests include development, implementation, and evaluation of active, contextual, student-centered curriculum and physical and computer-based biomolecular modeling activities to improve student learning, understanding, and engagement throughout the chemistry curriculum. She serves as a consultant to national and international audiences on the principles and practices of ability-based education, student assessment, and program/institutional assessment. For contact information, see http://www.alverno.edu/academics/academicdepartments/schoolofartssciences/chemistry/facultystaff/.

Anne McKee Ph.D. is a Senior Lecturer and Director of Educational Research and Innovation in "The School of Medical Education," King's College London. He has held administrative, research, teaching, and educational advisory roles at The University of East Anglia, The Open University, Milton Keynes, and Anglia Ruskin University. This experience has enabled him to work across institutions and professions, to develop collaborative working, build partnerships, and implement change in supporting learning in increasingly policy-driven and accountable climates globally. There are two underpinning lines of inquiry in his research. The first is to help professional and higher education anticipate and respond to change, much of it driven by policy and professional regulation which has created new educational demands. The second line of inquiry examines how higher education and professional education relate to each other as both transform. This involves studying the inter-relationships between policy, research and practice, and drawing implications for education in the professions, professional bodies, and diverse practitioners. He has experience in commissioning, directing and conducting research.

Danette W. McKinley Ph.D. is the Director of Research and Data Resources at the Foundation for Advancement of International Medical Education and Research (FAIMER). Dr. McKinley determines research priorities, defines scope, and proposes methodology for studies focused on understanding and promoting international medical education. She supports research activities related to the certification of graduates of international medical programs and plays a key role in the development of the Educational Commission for Foreign Medical Graduates (ECFMG) research agenda. Her interests include educational research methodology and assessment. She concentrates her efforts on the development of research programs on international medical education and the migration of health care workers.

John J. Norcini Ph.D. is the President and CEO of the Foundation for Advancement of International Medical Education and Research (FAIMER®). FAIMER has an active research program on international health professions education, databases of recognized medical schools, and accrediting bodies around the world, global fellowship programs for faculty from health professions schools, and in conjunction with Keele University, a Master's degree in Health Professions Education: Accreditation and Assessment. For the 25 years before joining the foundation, Dr. Norcini held a number of senior positions at the American Board of Internal Medicine. His principal academic interest is in assessment and he has published extensively, lectured and taught in more than 40 countries, and is on the editorial boards of several peer-reviewed journals in health professions education. He is an honorary Fellow of the Royal College of General Practitioners and the Academy of Medical Educators and has received numerous awards including the Karolinska Prize for Research in Medical Education.

Christopher O'Neal Ph.D. has worked in faculty development at the University of Michigan, the University of California, Irvine, and most recently as Director of Faculty Development at the David Geffen School of Medicine at UCLA. His research interests include problem-based learning, critical thinking, and multicultural issues in higher education. He is currently managing partner of a private venture in southern California.

Dan Osterweil M.D., F.A.C.P., M.sc. Ed., C.M.D. is the Vice President and Medical Director of SCAN Health Plan and a Professor of Medicine at UCLA and completed his geriatrics fellowship at UCLA. He is the Emeritus editor of the Journal of the American Medical Directors Association (JAMDA), which he founded. He is a member of the editorial board of Caring for the Ages. Dr. Osterweil co-authored two editions of*Medical Care in the Nursing Home*, is the co-editor of Comprehensive Geriatric Assessment, and has published over 60 articles in peer-reviewed journals. His areas of expertise include cognitive and functional assessment, management of dementia, continuous quality improvement in the nursing home, planning and implementation of work processes in the nursing home, in-depth knowledge of nursing home state and federal regulations, and geriatric practice innovations. Dr. Osterweil is the Director of a UCLA training program "Leadership and Management in Geriatrics" (LMG) and is Associate Director of the Multi Campus Program in Geriatrics and Gerontology at UCLA (MPGMG). Dr. Osterweil served as geriatric consultant to SCAN for many years prior to joining SCAN full time. Of his many duties at SCAN, Dr. Osterweil leads SCAN's senior-focused Healthcheck Assessment Center operations.

Yoon Soo Park Ph.D., is an Assistant Professor in the Department of Medical Education, College of Medicine, University of Illinois at Chicago. His areas of research interests are in assessment systems in medical education, psychometrics (item response theory and latent class models), and statistical modeling of psychological and social processes in medicine and public health. He also collaborates with an interdisciplinary team of methodologists and applied researchers to integrate innovative statistical techniques to refine measurement of learner performance and effectiveness in treatments. For contact information, see http://chicago.medicine.uic.edu/departments___programs/departments/meded/dme_faculty_staff/yspark2/.

Margaret Rauschenberger M.S.N., R.N., C.C.H.P. is a Professor of Nursing and Interim Dean of the JoAnn McGrath School of Nursing at Alverno College. Her research interests include retention strategies for diverse nursing students and the mental health needs of incarcerated adolescents. She has consulted nationally and internationally on educational outcomes development and student assessment and has presented extensively on workplace issues and culture. For contact information, see http://www.alverno.edu/academics/academicdepartments/joannmcgrathschoolofnursing/meetourfaculty/.

Scott Reeves Ph.D. is a social scientist who has been undertaking health professions education and health services research for over 20 years. He is Professor in Interprofessional Research, Faculty of Health, Social Care and Education, Kingston University and St George's, University of London and Editor-in-Chief, Journal of Interprofessional Care. He has spent the past decade leading interprofessional research in Canada and also in the United States. His main interests are focused on developing conceptual, empirical, and theoretical knowledge to inform the design and implementation of interprofessional education and practice activities. He has published numerous peer-reviewed papers, books, chapters, editorials, and monographs. Many of his publications have been translated from English into other languages including French, Spanish, Portuguese, Japanese, Norwegian, and Russian.

Joseph Rencic M.D. is an Associate Professor of Medicine and an Associate Program Director. In 2004, he started his career at Tufts Medical Center as a primary care practice in the division of general internal medicine. His research interests are in clinical reasoning and teaching residents to teach. He served as co-editor for a recently published book,*Teaching Clinical Reasoning*, for the American College of Physicians Teaching Series and has co-authored several book chapters on clinical reasoning assessment. Dr. Rencic received his undergraduate degree in biology from Georgetown University and his medical school degree from the University of Pennsylvania School of Medicine. He did his internship and residency at the Hospital of the University of Pennsylvania and stayed on as chief resident for the year following his residency.

Douglas R. Ripkey has been with the National Conference of Bar Examiners in Madison, WI for the past 10 years, currently holding the position of Deputy Director of Testing. He previously spent more than a decade at the National Board of Medical Examiners holding positions in both the Test Development and Psychometrics units with responsibilities for creating, scoring, and conducting operational and validation research on medical licensure and certification examinations. His research interests span the educational assessment spectrum, but he focuses currently on trends in legal licensure, measurement methodology, and test construction.

Mark Russell is a Professor of Learning and Assessment in Higher Education. He is the Director of Technology Enhanced Learning and the Head of the Centre for Technology Enhanced Learning at King's College London. Mark is an engineer by background and has taught in higher education since 1996. During this time Mark won the UK Times Higher Education e-tutor of the year (2003), became a National Teaching Fellow in 2005 and has led numerous institution-wide projects in relation to assessment and technology-enhanced learning. Mark has interests and expertise in assessment, curriculum design, and the thoughtful use of technology-enhanced learning.

William H. Rickards Ph.D. has over 20 years of experience as a Researcher in Educational Research and Evaluation at Alverno College. With a primary focus on program evaluation in higher education, he contributed to Alverno's landmark study of student learning and development, Learning That Lasts (Mentkowski and Associates 2000) and has continued evaluative studies of student learning and faculty practice in teacher education, nursing education, and engineering education. For several years, he contributed to studies of reflective learning and self-assessment in Alverno's diagnostic digital portfolio, editing a special Journal of General Education issue on electronic portfolios. He has also been involved as an external evaluator for an NEA Foundation project on urban schools, most recently studying the implementation of action research fellowships for Milwaukee teachers; he developed an experimental, graphic novel format for the evaluation report—Confessions of an Action Researcher: City Schools Confidential—to further study the relationships between practice-based approaches to action research (located in a school district) and academic approaches located in masters level education.

Brian S. Simmons B.Sc. (Hons), B.M., M.M.Ed., F.R.C.P.C. is a Clinician Educator, Associate Professor, Department of Paediatrics. Faculty of Medicine. University of Toronto and Neonatologist in the Division of Newborn Medicine at Sunnybrook Health Sciences Centre. Currently, Academic Director Standardized Patient Program University of Toronto. Chief Examiner Integrated OSCE, third year medical students University of Toronto and Deputy Registrar Medical Council of Canada (Toronto site.). Other Roles: Co-Chair of assessment in the neonatal perinatal medicine (NPM) program UT, Past Chair board of examiners NPM for Royal College of Physicians and Surgeons of Canada (RC). Assessment committee RC, examiner Paediatric OSCE RC. Development/implementation of an Interprofessional OSCE. Chair Awards committee Canadian association of Medical education (CAME). Past Roles: Faculty lead assessment in Interprofessional Education. The development/implementation/evaluation and assessment of the national OSCE in NPM. Scholarly Interests include the role of live simulation and Assessment. Integration of assessment methodologies related to performance/competence (OSCE). Assessment of learning in teams. Neonatal stabilization programs.

Personal favorite quote: Not everything that can be counted counts and not everything that counts can be counted—*Albert Einstein*.

Kelly Talley has worked in higher education for 12 years. She is the Director of the Assessment Center, Coordinator of the Diagnostic Digital Portfolio, and a member of the Council for Student Assessment at Alverno College. Kelly was a Member of the Design team for the patented DDP, and works with faculty and students to promote effective use of the DDP to support Alverno's ability-based curriculum. She has presented at a number of conferences on the Diagnostic Digital Portfolio

Ara Tekian Ph.D., M.H.P.E. is a Professor at the Department of Medical Education (DME), and the Associate Dean for the Office of International Education at the College of Medicine, the University of Illinois at Chicago (UIC). He joined DME in 1992, and is involved in both teaching courses offered in the Master of Health Professions Education (MHPE) program and advising graduate students. Dr. Tekian is an internationally recognized scholar and a leader in health professions education. He has organized and conducted over 250 workshops in more than 45 countries and 60 cities. His consultations and workshops have focused on curriculum development, assessment, program evaluation, and patient safety. He has received numerous honors and awards including the ASME (Association for the study of Medical Education) Gold Medal Award (2012), and the most revered Lifetime Achievement Award by the Armenian American Medical Society (2014). He has served as the President of the Division of Education in the Professions of the American Educational Research Association (AERA) from 2009 to 2012. His scholarship in health professions education is reflection in publications in the premiere medical education journals. He is the senior author of the book "*Innovative Simulations for Assessing Professional Competence: From Paper-and-Pencil to Virtual Reality*" published in 1999.

Susan J. Wagner B.Sc. (SPA) M.Sc. (CD) Reg. CASLPO S-LP(C) Associate Professor, Teaching Stream is the Coordinator of Graduate Studies and Senior Coordinator of Clinical Education in the Department of Speech-Language Pathology, Faculty of Medicine, University of Toronto (UT). She has provided leadership and program development in clinical education and been recognized nationally for mentorship in this area. She has also taught courses on principles of clinical practice and was integrally involved in planning and implementing an innovative curriculum for the clinical speech-language pathology Master of Health Science degree. Susan was the inaugural Faculty Lead—Curriculum from 2007 to 2012 at the Centre for Interprofessional Education (IPE), UT where she and her colleagues led the development and implementation of the requisite IPE curriculum for 11 health science programs that began in 2009. This involved creating and integrating IPE core competencies, learning activities and the points for interprofessional education system (PIPEs). Susan also played a fundamental role in the assessment, evaluation and faculty leadership components of the curriculum. She has been an investigator on a variety of research projects including development of an interprofessional objective structured clinical examination (iOSCE) and of IPE cases. The Susan J. Wagner Student Leadership Award in Interprofessional Education was named in her honor and she received the 2008–2009 University of Toronto Inaugural Award of Merit for Outstanding Leadership in Advancing Interprofessional Education through the Centre for IPE. Building on her interest and experience in competencies, she is co-chairing a committee to develop a national clinical education competency-based education tool in speech-language pathology and audiology. As well, she is conducting a research project on the development of milestones and entrustable professional activities (EPAs) for IPE. Susan has a keen interest in continuing professional development and faculty development and has given workshops on clinical education, interprofessional education (IPE) and dealing with conflict to health science professionals nationally and internationally. For publication and contact information, see https://www.researchgate.net/profile/Susan_Wagner/publications.

Noreen M. Webb Ph.D. is a Professor of Social Research Methodology in the Graduate School of Education and Information Studies at the University of California, Los Angeles. Her research spans domains in learning and instruction, especially the measurement and study of teaching and learning processes and performance of individuals and groups in mathematics and science classrooms, and measurement topics in generalizability theory. In her work related to educational measurement, she has produced numerous pedagogical pieces, has carried out a wide variety of empirical generalizability studies in education and psychology, and has co-authored a book on generalizability theory (*Generalizability Theory: A Primer*, with Richard Shavelson).

Michele Russell-Westhead Ed.D., M.Sc., P.F.H.E.A. is a Professor of Clinical Education working at King's College London, Northumbria University and Pearson College London where she is the Vice Principal for Education and Research. Her areas of research interest relate primarily to educational innovation, clinical simulation, industry-engaged education, employability, and professional issues in the health disciplines. She has an international profile as a curriculum innovator and academic leader in education.

Chapter 1
Introduction

Marcia Mentkowski and Paul F. Wimmers

Abstract Professional schools have some common goals, such as developing an identity as a professional person and encouraging lifelong learning. Yet they vary greatly in their characteristics and purposes. Educating professionals requires formal education. Yet, emphasizing formal and informal learning becomes essential with the rise of specialization and technology. Learning in the profession is best understood as a process embedded in social relationships and social practices, and other professionals, clients, learners, patients, and citizens participate in these relationships and practices over time and across settings.

Professional schools have some common goals, such as developing an identity as a professional person and encouraging lifelong learning. Yet they vary greatly in their characteristics and purposes. Educating professionals requires formal education. Yet, emphasizing formal and informal learning becomes essential with the rise of specialization and technology. Learning in the profession is best understood as a process embedded in social relationships and social practices, and other professionals, clients, learners, patients, and citizens participate in these relationships and practices over time and across settings (Curry and Wergin 1993; Peck et al. 2010; Wenger 1998).

The purpose of this book is to address how complex, learned abilities and capacities are assessed within and across different disciplines and professions. This goal, we think, can move the process of performance assessment-for-learning to the next level. We as editors are coping with the increasing complexity for finding the collaboration, teamwork, and resources in professional education. Thus we are committed to assessing competence in performance assessment as but *one strategy*. We as editors also believe that communication and collaboration among individuals of different professions is becoming even more challenging. This is obvious on the work floor of a large engineering and construction project. In managing complex

M. Mentkowski (✉)
Alverno College, Milwaukee, USA
e-mail: marcia.mentkowski@alverno.edu

P.F. Wimmers
David Geffen School of Medicine, UCLA, Los Angeles, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_1

rescue operations, launching a space shuttle, or producing big-budget movies, complexity reigns. Organizational environments require multiple areas of expertise and expect to develop the ability to work effectively with diverse stakeholders. Professionals deal with situations that have grown considerably more frustrating.

Graduates are expected to integrate their abilities and capacities, competencies and attributes with the content of their particular disciplines and professions, and also to engage in collaboration with other professional fields. In contrast to other theoretical and research-oriented higher education schools, professional schools often have well-defined performance outcomes tied to each profession which can be observed relatively directly. This heritage is still fundamental to the concept of occupational professionalism. Indeed, administrators, educators, and educational researchers could best engage in crucial systems thinking to engage with diverse systems. In fact, professional lives are becoming increasingly demanding, with even further specialization needed in professional development. Specialists and sub-specialists, who represent a narrower field of study within a discipline, have created their own unique definitions, acronyms, and terms. Yet the need for engagement in learning and human interactions increases in a global world.

Professional schools are now expected to graduate students who can evaluate their own work, continuously improve their performance, engage in lifelong learning, and also work at cross-professional problems. One issue is that all too often, interprofessional collaboration is thwarted by communication failures that take place at the boundaries between professions, organizations, and other groups (Edmondson 2012). Professional schools may lag in adapting to these new challenges. Nevertheless, rising expectations characterize learning in the workplace. This goes beyond learning the internal dynamics of homogeneous teamwork. Professional teaming goes across boundaries and has to focus on building bridges outside the team. Professional competence, an elusive term which is re-emerging as an approach to develop rigorous and broadly prepared graduates who are able to develop positive relationships with key stakeholders in and outside the organization they represent (Edmondson 2012; Parker 1994).

Situations that professionals deal with have grown considerably. Graduates not only are expected to integrate their abilities and capacities, competencies, and attributes with the content of their particular disciplines and professions, but to engage other professions in collaborative work. Are graduates able to integrate, apply, adapt, and transfer their capabilities across schools of thought, namely, the disciplines? And even to a degree, across professions? Administrators, educators, and educational researchers face new challenges, even when there are assessment strategies and technologies that offer support. Cross-professional comparisons can be at cross-purposes, in part because the nature of the work and its aims can differ so dramatically.

The authors of this book take up these issues as they apply new forms of assessment, along with definitions of terms. Performance is defined as:

> By performance we denote an individual's discretionary and dynamic action that effectively meets some contextually conditioned standard of excellence. Such multidimensional performance goes beyond technical or narrowly specified task performance. Performance entails the whole dynamic nexus of the individual's intentions, thought, feelings, and

> construals in a dynamic line of action and his or her entanglement in an evolving situation and its broader context. Such a context may be within or across work, family, civic, or other settings (Rogers et al. 2006).

The authors go on to define holistic development:

> By holistic development, we denote the overall direction of dispositional growth in the person's broadly integrated way of making meaning and commitments in moral, interpersonal, epistemological, and personal realms (Rogers et al. 2006).

Examples of issues common across professional schools include the transition from theory to practice and the impact on assessment, multiple definitions of competencies and their relationship to professional education, and increasing demands from stakeholders and accrediting groups. Other issues that cross professions include collaboration across professional groups in searching for solutions to professional problems. These problems include stretching to broaden one's understanding without losing sight of one's discipline. To resolve these issues, some educators are turning to assessment of role performances as one strategy for faculties to improve how they teach and for students to improve how they learn. Engagement with both faculty and students is essential in these kinds of complicated work settings across the professions.

To support this kind of learning, the authors review the development of constructs that cross disciplines and professions such as critical thinking, clinical reasoning, and problem solving. They also review what it takes for a faculty to develop competence in assessment, such as reliably judging students' work in relation to criteria from multiple sources. This implies that issues of quality in assessment and measurement are central to educational researchers' own capabilities. The professional continues to study following graduation, during further training, as a practitioner in the workplace, and throughout his or her life as a professional.

In medical care, for example, multidisciplinary teams are increasingly used for diagnosis and discussion of complicated treatment options and their outcomes. The opinions of an individual physician are making a place for higher order group decisions. An oncologist with a pancreatic cancer patient has to work together with surgeons, radiologists, palliative care physicians, nurses, dietitians, and hospital administrators. A psychiatrist involved in the assessment of an abuse case may work with professionals from other disciplines such as protective services or civil or criminal justice specialists. Teams are as diverse as the communities they serve. Focusing on improving coordination and communication between departments is of life importance for the future of health care (Frenk et al. 2010; Lamb et al. 2011; Ruhstaller et al. 2006; Tattersall 2006).

When the reliance on teams in organizations increases, team training and the evaluation of team performance becomes more important. In many situations, a well-functioning team can accomplish more than the sum of its individual persons (Doyle et al. 2013). Evaluation and assessment of team performance should focus on both, the performance of an individual in a team and the performance of the team as a whole. Team training starts in professional schools and many disciplines make use of simulated training for teams (Webb 1980).

1.1 The Nature of Interdisciplinary and Cross-professional Work

How well does our audience communicate with professionals from other disciplines and professions? What is the level of understanding roles and responsibilities of various team members? In the demanding context of a team, are graduates able to evaluate their own ideas in preparation to submitting them to a problem-solving team, for example, using the attributes that characterize a successful collaborator? Working together in a team with professionals across disciplines requires different skills and abilities than working as a team with your close colleagues who often represent a similar mental frame (See Edmondson 2012; Mintzberg 2004; Parker 1994).

Interaction among professionals in a multidisciplinary environment can take different forms and often we find that different descriptions are used in the literature: (1) multidisciplinary; (2) cross-disciplinary and cross-professional; (3) interdisciplinary; or (4) transdisciplinary.

For the purposes of this book, we define (1) *multidisciplinary* as developed breath of learning across disciplines that underlie one's profession, and (2) standing beside one's profession with enough depth that she/he understands which problems are facing other professions in one's sphere of influence. Our definition of *cross-professional* is: Interacting effectively across professions for identifying and resolving problems (e.g., problem-solving in an interprofessional team context that takes on meaning depending on the professions represented). Different professions may apply similar problem-solving techniques, but they solve very different problems and hence their performances are diverse.

Another goal for graduates is that they become able to *translate* their profession for other professionals without engaging in useless semantics and value conflicts. Rather, they can practice contributing to identifying, clarifying, and resolving some of the great problems of our time. These issues include arguing persuasively in community settings, and generating solutions in meetings by both brainstorming and useful critique (Isaksen 1998). Serving on panels of professionals to represent their own profession is a common occurrence. Yet, capturing and clarifying the ideas of another colleague to build on his or her ideas uses civil discourse, rather than engaging in competition promoted by "either, or" questions so familiar to us. Ultimately, we expect graduates to participate in developing sustainable policies in health care, for example. Do professional schools develop learners who can exercise these capabilities consistently and cooperatively by the time they graduate?

We also recognize that when the term, *performance assessment* arises, issues of instructional design, learning outcomes, and accreditations are also included in stakeholders' issues and demands. Along with educational policies across groups, states, and government policies, the editors adhere to the principle that a broad-based societal support system is critical for understanding the issues addressed.

1.2 Audience

This book is written not only for health care professionals, but also for faculty in the undergraduate professions, such as engineering, nursing, teaching, business and management, professional communications, and so on. Further, these undergraduate professions are likely to receive challenges from accrediting bodies for their performance assessments, and how they assess for abilities and capacities such as problem solving and critical thinking, valuing in decision-making, and a learner's ethical stance.

The role that administrators, educators, and educational researchers play as professionals in undergraduate and graduate professional schools already crosses disciplinary lines and professional boundaries. Both educators and researchers are often used to working in close collaboration with colleagues from other fields. For this reason, we believe the audience for this book already has wide experience engaging in border-crossing, and will be effective in engaging their colleagues across disciplines and professions to take up the issues related to this "great problem" of developing and assessing graduates with abilities and capacities, competencies and capabilities for these arenas. There are more cross-disciplinary and cross-professional problems that have been the subject of symposia by AERA Division I, Education in the Professions. This volume is built on one such symposium titled, *Clarifying Assessment and Measurement Issues across Disciplines and Professions,* organized by Danette McKinley. The symposium had nearly 90 participants, who stayed to define issues in small groups, organized around the following questions.

1. How might a competence-based education model with learning outcomes impact learning and assessment?
2. What are some challenges in assessing a student's learning outcomes across disciplines and professions?
3. What definitions in critical thinking capture differences in assessing this ability across the professions? This ability takes multiple forms in context. What criteria should be assessed? One approach is to respond with developmentally articulated levels.
4. What challenges do we face when building consensus definitions of professional constructs for developing and implementing performance assessments?

A second proposal Assessing Competence in Professional Performance across Disciplines and Professions and organized by Marcia Mentkowski, captured the issues that remained to be discussed from the initial symposium in 2013. In 2014, Anne McKee (King's College London) and EunMi Park (Johns Hopkins University) conducted the question and answer session and wrote up the questions for *Professions Education Research Quarterly.*

1.3 New Developments in Performance Assessment for Undergraduate Education

Educational researchers are familiar with performance assessment-for-learning. *Performance, or a set of performances*, is defined as what a learner does with what he/she knows, in context. This usually involves performance in a particular role, or across roles. Recall that *multidimensional performance* entails the whole dynamic nexus of the individual's intentions, thoughts, feelings, and construals in entanglement in an evolving situation and its broader context. Such a context may be within or across various roles in the professions (Rogers et al. 2006).

Another term often used is *authentic assessment*. This term refers to the kind of task (open-ended, constructed) and the quality of the relationship to performances that learners will need to demonstrate in their other positions post-graduation. Assessment of role performance directly related to performance after professional school is essential in three areas: (a) student assessment-for-learning; (b) program evaluation or assessment-for-improvement; and (c) institution-wide assessment that may be primarily for improvement, but is often used for demonstrating accountability to stakeholders.

We offer the approach of *performance assessment* defined as judgment by self, peers, faculty, or other assessors and mentors in relation to multiple, diverse criteria or standards derived from practices which are determined by the contexts of practice. Performance assessment often includes self-reflection and self assessment, namely, observing, interpreting or analyzing, and judging ones performance. This learning process is for deepening ones learning and competence, and planning for further learning. In this case the person is the *agent* and not the *object* of assessment. The role of feedback is crucial. Feedback is *accurate, conceptual, diagnostic, and prescriptive.* The role of feedback in teaching cannot be underestimated (Alverno College Faculty 1979/1974; 2015).

These ideas are a partial solution to The National Academies' National Research Council study of test-based incentive constructs. The study found that few learning benefits accrued from high-stakes exams for faculty or students (Hout, Elliott, and Committee on Incentives and Test-Based Accountability in Public Education 2011). This K-12 pattern repeats in higher education as well. Fortunately, there are examples of performance assessment at every level of practice.

One example is the Collegiate Learning Assessment, which uses performance tasks primarily for institution-wide assessment for accountability, and less often for curriculum improvement. It is used to assess broad competences such as communication, critical thinking, and problem solving—usually in the role of "learner." It is used across undergraduate disciplines and professions. This instrument is also currently being developed for international use, to potentially replace or augment other instruments used for comparing student performance across countries. Using this kind of comparison may be useful for some purposes, but it is not as useful for finely tuning a curriculum where the purpose is developing highly competent and capable graduates, providing them with feedback, asking them to engage in self assessment so they learn to evaluate their own work, or to assist them to figure out what is next on their learning agenda.

A second example of performance assessment is the American Association of Colleges and Universities project. The association received funding for three projects from the U.S. Department of Education, one of which was called Valid Assessment of Learning in Undergraduate Education (VALUE). While the focus was program evaluation, the project consisted of enrolling schools that had experimented with electronic portfolios. They reviewed other colleges' eportfolios, that is, collections of student work. These performances, where ones performance is defined as a generative sample or samples from a collection of student work, were reviewed by various cross-institution and cross-disciplinary teams who then developed rubrics for various student learning outcomes that most undergraduate colleges have in common (intellectual and practical skills, ethical decision-making, integrative, applied, and adaptive learning, and so on). This kind of cross-institution assessment with dimensions and criteria that describe learning outcomes is an important step forward in program evaluation, and provides examples for the kinds of competencies that are discussed in this volume.

This learning process includes reflective learning on one's own abilities, self assessment, and feedback. Learners also engage in integrative, applied, and adaptive learning, and what about transfer of learning or transformative learning? These are fairly new constructs for the audience for this book. While faculty may not have a particular mindset for these constructs, we expect educational researchers to be ready to explain them.

Finally, performance assessments of abilities and capacities, competencies, attributes, and dispositions are consistently used in undergraduate and graduate professional schools *to infer competencies from a role performance or set of performances.* Faculty at these schools often hold themselves to a model of abilities, capacities, or competencies. Thus, the quality of measurement and assessment are equally important to this book.

1.4 Concluding Argument

As editors, we are aware that the role of fostering interdisciplinary education without a committed faculty who identifies with their role as teacher and assessor is a tall order. We do not plan to attempt such goals alone, as educational researchers. But we believe that without the capabilities of cross-professional competencies and their transfer across settings where graduates confront other professional problems, professional school graduates will not meet the demands of today's complex situations and settings. True, professionals have different knowledge systems and values, and one cannot ask them to know everything. Yet we believe graduating students who use, adapt, and transfer learning outcomes or competences to fast-moving settings will be more effective. These settings will change dramatically from the time professionals enter their life's work until they leave it, but they will be more humanistic and ethical professionals.

We as administrators, educators and researchers take responsibility for our students' learning seriously. Developing cross-professional competence is not the only answer. It is only one part of what graduates will face. It is but one way to support our graduates in retaining their fervor and idealism in the face of the often excessively brutal indifference and insensitivity occasioned by modern work. Collaborating professionals who share similar problems may be able to offer life-long support for those who work beside them. We believe that at this point in history, many professionals face resistance, and even isolation. In sum, the authors define abilities and competences, a range of performance assessments, and the consequences of assessment for society.

New insights are needed about the nature of professions education: (1) ways professionals interact and work together in teams with other professionals in multi-professional settings and (2) consequences for learners for performance assessments in both undergraduate and graduate professions. Perhaps implicitly, authors are asking "who cares?" and answering "we do."

References

Alverno College Faculty. (1979/1994). Student assessment-as-learning at Alverno College. Milwaukee, WI: Alverno College Institute (Original work published 1979, revised 1985 and 1994).

Alverno-College-Faculty. (2015). *Feedback IS teaching*. Milwaukee, WI: Alverno College Institute.

Curry, L., & Wergin, J. F. (Eds.). (1993). *Educating professionals: Responding to new expectations for competence and accountability*. San Francisco: Jossey-Brass.

Doyle, C., Wilkerson, L., & Wimmers, P. F. (2013). Clinical clerkship timing revisited: Support for non-uniform sequencing. *Medical Teacher, 35*(7), 586–590. doi:10.3109/0142159X.2013.778393

Edmondson, A. C. (2012). *Teaming: How organizations learn, innovate, and compete in the knowledge economy*. San Francisco: Jossey-Bass.

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., et al. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet, 376*(9756), 1923–1958.

Hout M., & Elliott, S. W. (Eds.). (2011). Committee on incentives in test-based accountability in public education. In: *Incentives and test-based accountability in education*. Washington, DC: National Research Council.

Isaksen, S. G. (1998). *A review of brainstorming research: Six critical issues for enquiry (Monograph #302).* Buffalo, NY: Creative Problem Solving Group-Buffalo.

Lamb, B. W., Sevdalis, N., Arora, S., Pinto, A., Vincent, C., & Green, J. S. (2011). Teamwork and team decision-making at multidisciplinary cancer conferences: barriers, facilitators, and opportunities for improvement. *World Journal of Surgery, 35*(9), 1970–1976.

Mintzberg, H. (2004). *Managers, not MBAs: A hard look at the soft practice of managing and management development*. Berrett-Koehler.

Parker, G. (1994). *Cross-functional Teams*. San Francisco: Jossey-Bass.

Peck, C., Gallucci, C., & Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education, 61*, 451–463.

Rogers, G., Mentkowski, M., & Reisetter Hart, J. (2006). Adult holistic development and multidimensional performance. In C. Hoare (Ed.), *Handbook of adult development and learning*. New York: Oxford University Press.

Ruhstaller, T., Roe, H., Thurlimann, B., & Nicoll, J. J. (2006). The multidisciplinary meeting: An indispensable aid to communication between different specialities. *European Journal Of Cancer (Oxford, England: 1990), 42*(15), 2459–2462.

Tattersall, M. H. N. (2006). Multidisciplinary team meetings: where is the value? *The Lancet Oncology, 7*(11), 886–888.

Webb, N. M. (1980). Group process: The key to learning in groups. *New directions in the methodology of social and behavioral research, 6*, 77–87.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Chapter 2
Conceptual Elements for Performance Assessment for Faculty and Student Learning

Marcia Mentkowski, Mary E. Diez, Dena Lieberman, Desiree Pointer Mace, Margaret Rauschenberger and Jeana Abromeit

Abstract This conceptual chapter clarifies elements for performance assessment that hold promise for designing performance assessments, including capstone and portfolio assessments. Elements were originally determined by Alverno College faculty from their practice in 1973 and combined with an internal and external literature review of relevant theoretical frameworks across time. This literature review included many early citations of such concepts as active learning, self-reflection and self-monitoring, assessment and judgment in relation to criteria, and the role of samples of performance in assessment. For this chapter, citations from literature external to the College and Alverno literature have been recently reviewed and illuminated for the following elements of performance assessment, also articulated as learning processes, transformative learning cycles, and learning outcomes. (1) Self-reflection on learning own abilities; (2) Self assessing performance and using feedback to improve it over time; (3) Learners developing metacognitive performance; (4) Learners developing professional expertise; and (5) Learners developing identity as a self-sustained and unique learner, contributor, and professional.

M. Mentkowski (✉)
Professor Emerita, Psychology, Alverno College, Milwaukee, WI, USA
e-mail: marcia.mentkowski@alverno.edu

M.E. Diez
Professor Emerita, Education, Alverno College, Milwaukee, WI, USA

D. Lieberman
Business and Management, Alverno College, Milwaukee, WI, USA

D. Pointer Mace
Education, Alverno College, Milwaukee, WI, USA

M. Rauschenberger
Nursing, Alverno College, Milwaukee, WI, USA

J. Abromeit
Sociology, Arts and Humanities, Alverno College, Milwaukee, WI, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_2

Takeaways

- Self-reflection on learning one's own abilities is an early element that the faculty worked at capturing through their use of self assessment.
- Self assessment leading to self-confidence and self-efficacy emerged as central to learning.
- Integration of subject matter and learned abilities, with subsequent adaptation and transfer of learning outcomes to unscripted settings is a predominant goal for education.
- Learners developing professional expertise related to self assessing role performance.
- Learners developing an identity as a self-sustained and unique learner contributor, and professional, is one of the more challenging learning outcomes for educating professionals.

2.1 Introduction

Most authors agree that determining learning outcomes as an integration of subject matter and learned abilities, an essential element for performance assessment that leads learners to become competent professionals, begins the faculty learning process, because most faculty begin this task to exercise their professional responsibility and apply their expertise. Further, faculty members often start by making student learning outcomes more explicit. For example, as Alverno College faculty worked to accomplish this, they brought theoretical frameworks underlying their disciplines to the task, thinking them through and deciding what content and abilities should be learned in their courses. Loacker and Palola described this process as early as 1981, and they gave examples of institutions that were joining them in carrying out this process.

2.2 Early Literature: Sources of Evidence for Early Practice

Thus, when Alverno faculty derived elements for performance assessment from their practice and the early literature, student opportunities for (1) self-reflection on learning one's own abilities was an early element that the faculty worked at capturing through their use of self assessment, where the "self" was the *agent* rather than the *object* of assessment. They videotaped each student on the first day of college as each gave a speech. Faculty members then engaged students in self assessing their performance, rather than their person.

Literature on performance assessment in higher education was lacking in the early 1970s, when Alverno was developing its student assessment as learning process. Alverno College had contracted the Educational Testing Service (ETS) to develop measures of their abilities as learning outcomes (or competences as they were then called) (Alverno abilities: Communication, Analysis, Problem Solving, Social Interaction, Valuing in Decision-Making, Effective Citizenship, Developing a Global Perspective, and Aesthetic Engagement). However, Alverno abilities were then and are now defined as complex capabilities of the person that can be taught, learned, assessed, observed in performance, and continually rethought and refined (Alverno College Faculty 1973/2010; Mentkowski 2006; Mentkowski and Sharkey 2011).

After a year of tryouts, ETS suggested that measures of these learning outcomes or competences would need to be faculty designed, because a technology for those kinds of assessments was not yet available at ETS. However, they noted that Lois Crooks of ETS was working to develop and validate the *In-Basket* (1973), an assessment she had designed to elicit performances. Alverno faculty adapted this measure for a part of their first *Mid-Program Performance Assessment of General Education* and studied it for establishing validity.

By the 1980s, several pieces had been published that opened the door to better understanding of students and reflection on their learning, as well as self assessment and the performance assessment process and its validity. Each of these chapters and articles contributed to the early literature available to Alverno faculty (Alverno College Faculty 1979, revised 1985 and 1994; Anastasi 1980; Boud et al. 1985; Friedman and Mentkowski 1980; Friedman et al. 1980, 1982; Loacker et al. 1986; Loacker and Jensen 1988; Loacker and Mentkowski 1982; Marton et al. 1984; Mentkowski and Doherty 1984; Mentkowski and Loacker 1985). This literature also stimulated the student and program assessment movement in higher education (Ewell 1985).

Because self assessment was part of Alverno's student as learning performance assessment process, a second element, (2) self assessment leading to self-confidence and self-efficacy emerged as central to learning (Alverno College Faculty 1979/1994; Bandura 1986). Thus, a third element began to emerge from faculty who were eliciting student samples of self assessment of their own performances. These samples, among other factors such as their teaching and learning in their disciplines, were helpful for faculty because they could see and imagine how students were experiencing their learning. This reinforced faculty members to continue to develop learning experiences for integrating subject matter and learned abilities (Alverno College Faculty 1976/2005; Anastasi 1980). These student samples became a further source of criteria and standards at Alverno.

Student samples were also being used for similar purposes, integration of coursework and abilities, at that time in some schools in K-12 education and in postgraduate education in the professions Alverno invited participants from across the educational spectrum: Alverno College, Milwaukee, Wisconsin; Bloomfield Hills Model High School, Bloomfield Hills, Michigan; Central Missouri State University, Warrensburg, Missouri; Clayton College and State University, Morrow, Georgia; Purdue University School of Pharmacy and Pharmacal Sciences, West

Lafayette, Indiana; South Division High School, Milwaukee, Wisconsin; Township High School District 214, Arlington Heights, Illinois; University of New Mexico School of Medicine, Albuquerque, New Mexico; and University of Wisconsin Medical School, Madison, Wisconsin. (Consortium for the Improvement of Teaching, Learning and Assessment (1992/June). (Funded by a grant to Alverno College from W. K. Kellogg Foundation (1989–1992).

From across the educational spectrum, faculty members began learning that integration of subject matter and learned abilities, with subsequent adaptation and transfer of *learning outcomes* (namely, integrated subject matter and learned abilities) to unscripted settings was a predominant goal for education, and consequently, for assessment (Alverno College Faculty 1979/1994; Boyatzis 1982; Loacker and Palola 1981; McClelland 1973). Gradually, element (5) professional role performances for developing identity as learners (Kegan 1982) and professionals (Argyris and Schön 1974; Schön 1973, 1983, 1987) emerged as well, especially given Schön's early conceptualization (1973) of such ideas as *reflective-practitioner, theory-in-use, reflection-in-action,* and *learning-in-use.* Schön's ideas also led to element (4) learners developing metacognitive performance.

2.3 Later Literature: Sources of Evidence for Current Practice

The later literature confirms the first two elements of performance assessment. To prompt *self-reflection*, faculty members across higher education may engage students in mentally creating narratives that are personally meaningful and reflect their values, relying on immediate experience in context (Baxter Magolda and King 2007). Students also listen to others' stories of similar events, or reflect on their own performance in situations (Bandura 1997; Boud and Walker 1998; Mentkowski et al. 2000). As faculty members listened to student stories, they began to notice particular patterns on how students think and learn, as Perry (1970, 1980) demonstrated over four decades ago. Alverno College confirmed these patterns with an independent, longitudinal study (Mentkowski et al. 1983).

To prompt *self assessment* ("self" is not the *object* but the *agent* of assessment), faculty members may engage students in deepening their learning by observing, analyzing and interpreting, and judging their own role performances in various situations in relation to criteria; and then encouraging learners to plan for further learning to see how to improve it. Central to this learning process is the role of an instructor, who provides *accurate, diagnostic, conceptual, and prescriptive feedback on performance* to students. If learners are to use feedback to improve their performance over time, faculty feedback is essential. Alverno learners developed their skill in learning to use feedback (see Mentkowski et al. 2000, Appendix H: Developing Perspectives of Students on Self Assessment, Using Feedback, Commitment to Improvement, and Role of Criteria, for beginning, intermediate, and advanced student perspectives, pp. 447–451). Georgine Loacker edited *Self*

Assessment at Alverno College, which provides a plethora of examples of instructors promoting self assessment and observing it in their students' learning (Alverno College Faculty 2000).

Peers and community professionals may also provide feedback and make judgments on standards of practice. A professional school graduate faculty may consider these perspectives when judging whether and how students have met criteria and standards (Gruppen et al. 2000). Faculty members are often learning what has been effective and what instructional methods needed changing. Instructors may model this analytic learning process by providing examples of where a student succeeded, and what needs work (Alverno College Faculty 2015).

To prompt a third element of performance assessment that involves integrating subject matter and learned abilities, faculty members may stimulate students' integration of conceptual frameworks and intellectual and practical skills through studying and eliciting performance (American Association of Colleges and Universities 2011; Eraut 1994). Faculty members may also design pedagogically developmental assessments for students to practice integrating content and skills. When students also integrate constructs with learned abilities, they gradually become capable of demonstrating professional roles in their labs and internships. Faculty members have been learning how students build representations of knowledge integrated with abilities as observed in practice. Alverno faculty members have also been learning to make finer and more nuanced adjustments in what is pedagogically developmental.

Alverno students have learned to draw on their capacity for reasoning, for using metacognitive strategies as frameworks for connecting their learned abilities to actual performances through simulated performance assessments. These metacognitive strategies or learned abilities assist students to recognize patterns, and enable them to think while they are performing—including about disciplinary frameworks. Thus, learners are restructuring their knowledge so they can adapt it to a scripted setting as a prelude to adapting their knowledge to an unscripted setting, again via a performance assessment. Thus, thinking metacognitively across knowledge-rich environments becomes essential to learning that lasts (Mentkowski et al. 2000). This leads to a third element, *learners developing metacognitive performance.* To prompt a fourth element, *learners developing professional expertise,* faculty often develop examples that assist professional school students to also adapt their *learning how to perform*. Here, faculty members who have professional experience can tap their own understanding for what it takes for their students to develop a picture of a performance in an interactive setting (which describes most professional roles). Such a performance assessment requires a learner to carefully observe a setting for clues for "knowing what to do when I don't know what to do."

Another transformative learning cycle that Alverno students experience in an ability-based curriculum is named Self Assessing Role Performance (Mentkowski et al. 2000). Such a process is essential for adapting and then transferring learning outcomes, even though other authors may use different languages for the development of expertise (Ericsson et al. 2006; Sternberg 1998). Transfer of learning has been further substantiated by Boyatzis et al. (1995) in their management curriculum. Bransford et al. (2000) developed schema for generalizing transfer of learning. Hakel and Halpern (2005) argue that transfer can happen across physical, temporal, and conceptual space. Mestre (2005) edited a volume about transfer of learning from multidisciplinary perspectives. Both sources provide faculty with insights about what it means to transfer learning. Mestre also shows what it means to engage in cross-disciplinary and cross-professional learning and transfer.

2.4 Developing a Faculty-Designed *Mid-program Performance Assessment for General Education*

Faculty members may require students not only to practice but to demonstrate, adapt, and transfer *learned capacities*, defined as integration of learned abilities with patterns of performance, or dispositions of the person. Usually, faculty members design and construct performance situations that are unfamiliar to learners, so that faculty members can judge student capacities at integration and transfer (see Abromeit 2012; Mentkowski et al. 2012). (The first chapter, written for the proceedings of the Higher Learning Commission, and the second chapter prepared for an Education in the Professions: Division I Symposium, provided a description and analysis of Alverno's *Mid-Program Performance Assessment for General Education* that all current students complete between their second and third years at Alverno, usually about the time they enter their professional fields. The second chapter also described the validation process for this assessment.)

In the context of Alverno's faculty-designed *Mid-Program Performance Assessment for General Education,* students first complete a self-reflection and self assessment on their own. Following, trained faculty assessors who observe and assess performances from across the disciplines and professions, may continually learn that adaptation and transfer are an extension of integrating subject matter and learned abilities. This is because students must demonstrate their expertise through adapting their communication ability, integrated with their scientific reasoning and mathematics, along with their problem solving and quantitative reasoning abilities. They are required to demonstrate content integrated with abilities in an unfamiliar, but knowledge-rich environment in order to be successful. This summative performance assessment is combined with faculty assessor feedback and judgment of success or lack of success. Faculty assessors also draw out students by interacting with them on their ideas for further learning, whether or not a student has been successful on this summative assessment. Thus, faculty-designed performance assessments may become a source for reciprocal learning by both faculty and students. This is especially the case when a learner is unsuccessful, attends a reassessment workshop, and then is required to transfer her abilities to a different but related performance assessment.

As noted above, when students themselves began to view not only assessing their performance and using feedback to transform it, but also engaging in reflective learning, envisioning improved role performances, and then monitoring role performances in relation to criteria from diverse sources, they have experienced Self Assessing Role Performance, a transformative learning cycle. Here learners may then find that self assessment is a useful pathway to improvement across multiple, varied situations.

A fifth element, *learners developing an identity as a self-sustained and unique learner, contributor, and professional,* is one of the more challenging learning outcomes for educating professionals. Learners may take multiple paths, however unique, to engage their development of identity as a professional (Bebeau and Monson 2012; Gruppen et al. 2000; Mentkowski et al. 2000) This is a gradual learning process, but without faculty members who have been stimulating students to make connections, Alverno educational researchers have learned that not all learners may make them where they need to—in a professional context. Alverno faculty members have also learned that to stimulate this learning process, they need to provide developmentally appropriate examples that demonstrate relevance for the social learning of each professional role. Following such stimulation by professional school faculty, students may demonstrate the transformative learning cycle, Engaging Diverse Approaches, Views, and Activities. Most faculties observe that students learn independently before they learn interdependently. This usually requires some developmental restructuring of their thinking and reasoning, even as they begin to learn interdependently. Faculty members sense that this is a more sophisticated way of learning, because it is usually interactive. Learners may also strive for mutuality—learning to engage others as professionals, which faculty members often stimulate and simulate through internships and clinicals for students to develop learning outcomes (Mentkowski et al. 2000).

2.5 Development of Practice in Performance Assessment by Alverno Faculty Across Time

Over four decades, the Alverno College faculty has continually improved their practice in performance assessment with ongoing workshops conducted with faculty and academic staff via the Alverno Faculty Institute. These workshops occur semester-by-semester, and are usually conducted by the Council for Student Assessment. These workshops have consistently illuminated appropriate literature both external to the College and Alverno literature. Since 1976, its current Senior Scholar for Educational Research has maintained the literature review in the department of Educational Research and Evaluation (For these reviews, the reader may refer to Mentkowski and Loacker 1985; Mentkowski 1998; Fig. 2.1).

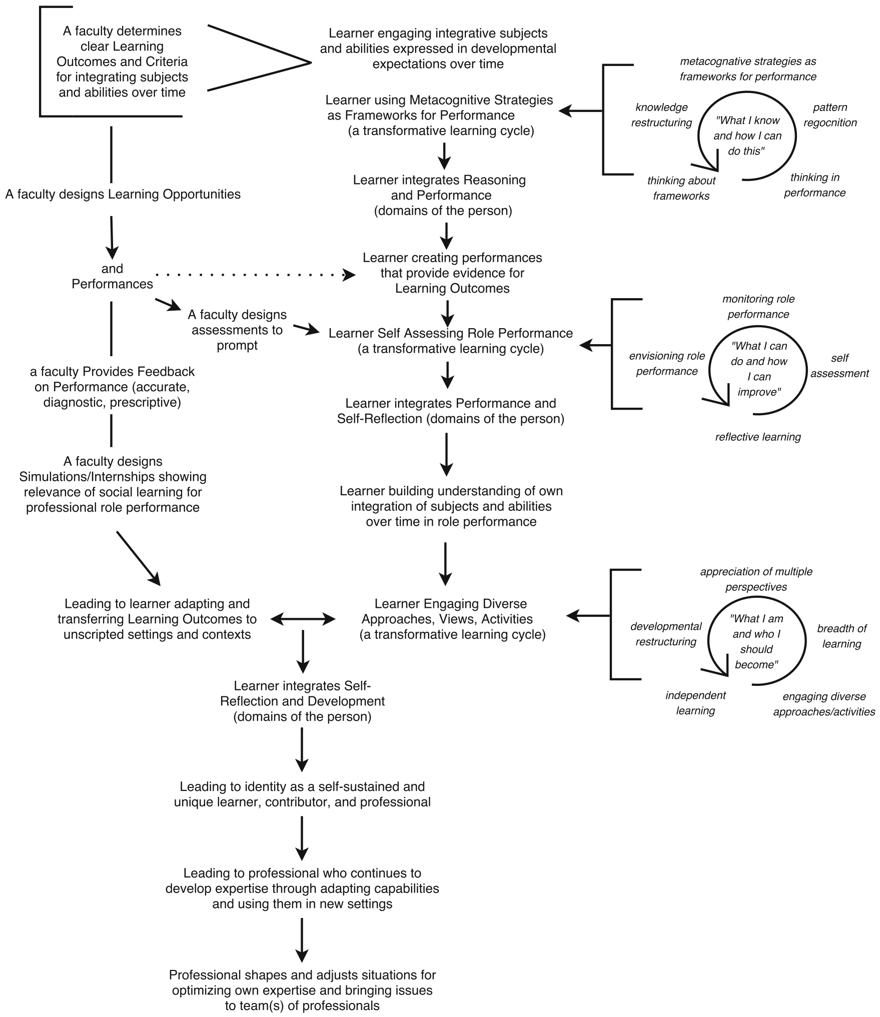

A faculty determines clear Learning Outcomes and Criteria for integrating subjects and abilities over time
Learner engaging integrative subjects and abilities expressed in developmental expectations over time
Learner using Metacognitive Strategies as Frameworks for Performance (a transformative learning cycle)
metacognative strategies as frameworks for performance
knowledge restructuring
"What I know and how I can do this"
pattern regocnition
thinking about frameworks
thinking in performance
A faculty designs Learning Opportunities
Learner integrates Reasoning and Performance (domains of the person)
and Performances
Learner creating performances that provide evidence for Learning Outcomes
A faculty designs assessments to prompt
Learner Self Assessing Role Performance (a transformative learning cycle)
monitoring role performance
envisioning role performance
"What I can do and how I can improve"
self assessment
reflective learning
a faculty Provides Feedback on Performance (accurate, diagnostic, prescriptive)
Learner integrates Performance and Self-Reflection (domains of the person)
A faculty designs Simulations/Internships showing relevance of social learning for professional role performance
Learner building understanding of own integration of subjects and abilities over time in role performance
appreciation of multiple perspectives
developmental restructuring
"What I am and who I should become"
breadth of learning
independent learning
engaging diverse approaches/activities
Leading to learner adapting and transferring Learning Outcomes to unscripted settings and contexts
Learner Engaging Diverse Approaches, Views, Activities (a transformative learning cycle)
Learner integrates Self-Reflection and Development (domains of the person)
Leading to identity as a self-sustained and unique learner, contributor, and professional
Leading to professional who continues to develop expertise through adapting capabilities and using them in new settings
Professional shapes and adjusts situations for optimizing own expertise and bringing issues to team(s) of professionals

Fig. 2.1 Relationships Among Faculty Designs, Integrative Learning, Transfer of Learning, and Identity Development Prompted by Elements of Performance Assessment

2.6 Alverno Assessment Center for Student Learning and Development: How It Began

The Alverno faculty has had the support of an assessment center throughout their work with performance assessment. Indeed, Alverno College was the first of any organization to use the assessment center method for the purpose of *student learning and development.* The assessment center method was adapted to include elements of the process, but the faculty learned to use this process for educational purposes such as: (a) Clear learning outcomes and criteria, (b) eliciting performance samples, (c) instructor and peer feedback on performance samples, (d) self assessment where the "self" is the *agent* rather than the *object* of assessment, (e) judgment of samples in relation to criteria by trained assessors, and (f) assessor and learner setting goals for further learning. In the case of a summative assessment, the goal of the assessor is to make judgments about strengths and areas to develop and discuss these with the learner, the role of the learner is to listen carefully to the feedback, and then the assessor and learner use the feedback in planning for further learning (Alverno College Faculty 1979/1994).

At that time, Dr. Douglas Bray at AT&T was using the assessment center method for *identification*, with *rejection* and *selection* of other business and management professionals, for their further training and advancement in the AT&T organization (Sister Joel Read, personal communication, March 15, 19, 26, 2013). William Byham and Douglas Bray cofounded *Development Dimensions International* in 1970. From 1969 to 1972, Alverno College had been generating its ability-based learning outcomes and its curriculum with its student assessment process for learning and development (Read and Sharkey 1985). In the academic year beginning fall 1972, an academic task force was charged by the faculty with "synthesizing faculty ideas into a blueprint for a curriculum" (Loacker et al. 2011). A group of four Alverno faculty members, Dr. Austin Doherty (Ph.D., Loyola University, Chicago, Psychology), Dr. Georgine Loacker (Ph.D., English Literature, University of Chicago); Jack Cooper (Masters in Music, Catholic University), and Dr. Brian Nedwek (Ph.D., University of Wisconsin, Milwaukee, Political Science), were invited by the faculty to research the learning outcomes that the corporate faculty had determined. In the process of her research, their assistant, Ms. Betsy Dalpes, found a description of an assessment center.

Sister Joel Read, Alverno's President, spent an entire day in 1971 calling each of the contacts that Ms. Dalpes had uncovered. Thus, she learned about the assessment center method at AT&T. In 1972–73, President Read and her faculty colleagues, Dr. Austin Doherty (professor of psychology) and Dr. Georgine Loacker (professor of English), traveled to New York's AT&T in lower Manhattan to find out more (Sister Joel Read, personal communication, March 15, 19, 26, 2013). They met with Joel Moses of AT&T. Moses suggested they visit Les Weinberger, an industrial psychologist and director of the AT&T Assessment Center in Milwaukee, Wisconsin. This they did, learning from Les Weinberger and later, Dr. Kelly Conrad, about this method used for *identifying*, *rejecting,* and *selecting* business and management professionals (Read, personal communication, March 15, 19, 26, 2013). Later, Weinberger and Kelly were loaned to Alverno through a community service agreement.

Loacker also traveled to the United States Military Academy (USMA) at West Point, NY to learn what the academy was doing related to identification, rejection, and selection of cadets. Loacker connected to USMA because Robert Greenleaf had traced the assessment center method, during an AT&T visit, to psychologist Henry Murray's successful assessments of performance for *identification, rejection,* and *selection* of effective professionals for spymaster service in World War II. In 1943, Murray was a member of the Office of Special Services. In 1956, at AT&T, Robert Greenleaf had read, “A Good Man is Hard to Find”, an article in *Fortune* on OSS assessment. Following, Greenleaf brought on Douglas Bray to design a new assessment for selecting managers, similar to OSS Assessment. Henry Murray was also involved in this new assessment design (Loacker et al. 2011).

Thus, Alverno’s Assessment Center for Student Learning and Development, the first in any organization, was born and is ongoing. The curriculum was institutionalized in 1973, and in 1976, Alverno instituted its program and institutional assessment process and its longitudinal studies, with a grant from the National Institute of Education at the U.S. Department of Education.

However, all performance assessments used in the assessment center are *faculty designed*, as evidenced by the two assessments detailed below, in order that they can not only be used for learning and development of students, but also faculty learning and development, so that faculty members may engage in curriculum improvement. We now move to additional modes of inquiry that test the elements against faculty practices, namely faculty-designed performance assessments. Other measures, especially those used in longitudinal studies, were administered by the department of Educational Research and Evaluation.

2.7 Additional Modes of Inquiry

Two analyses of assessment designs from practice at Alverno College were used for selecting one summative *faculty-designed* performance assessment from each of the professions of nursing and business and management. Each appeared to exemplify each of the five elements. This chapter concludes with an example of a *profession-designed* assessment of student teaching practice that appeared to exemplify three of the five elements of effective performance assessment.

2.7.1 Alverno College School of Nursing Capstone Performance Assessment and the Alverno College School of Nursing Professional Portfolio

Assuring competence of students in professional nursing practice is critically important in a nursing education program. Using theoretical frameworks from the discipline of nursing and the expected learned abilities along with the principles and standards inherent in professional nursing practice (American Nurses Association 2001, 2003) the Alverno School of Nursing faculty developed a set of program outcomes that describe the required student performance. The *Alverno College School of Nursing Capstone Performance Assessment* and the *Alverno College School of Nursing Professional Portfolio* were subsequently faculty designed to reflect the program outcomes.

Effective strategies for evaluating outcomes include both faculty judgment of performance and student awareness of level of expertise. While assessments and subsequent feedback to students can be formative or summative, assessing whether students have met program outcomes requires a summative assessment that includes student self-reflection and self assessment of performance, using feedback in order to transform it. The capstone assessment in the nursing program is designed as a mock employment interview for a new graduate nursing position in a hospital setting. The interviewers are volunteer assessors from the nursing community who have been trained in administering the assessment, using the criteria to make judgments, and providing feedback to students so they can improve their performance. Assessors are given a list of questions to ask the student in the interview so that interviews are consistent across students.

2.7.2 Alverno College Professional Nursing Portfolio[1]

In preparation for the assessment, students are given the assessment criteria for review, provided a workshop on interviewing skills, and faculty members have assessed their *Alverno College School of Nursing Professional Portfolio* for nursing criteria related to program outcomes. They are expected to bring and use their portfolios in the interview, and volunteer assessors from the nursing community base their observations, judgments, and feedback on the portfolio as well.

Topic areas. Nursing students are also given the following topic areas as possible themes for interview questions (Table 2.1).

- Cultural diversity,
- Ethical dilemmas,
- Leadership/delegation/prioritization,
- Use of theory in nursing practice,
- Problem solving in difficult situations,
- Conflict resolution,
- Participation as a team member,
- Future goals,
- Personal qualities and assets and areas to develop,
- Patient advocacy, and
- Professionalism.

[1] Section prepared by Margaret Rauschenberger, Professor of Nursing and Interim Dean, JoAnn McGrath School of Nursing, Alverno College, and Master of Science in Nursing (MSN), Registered Nurse, (RN), and Certified Correctional Health Professional (CCHP). She serves on the Curriculum Committee, Educational Policies Committee, and Problem Solving Department. Rauschenberger also maintains her practice as an advanced practice nurse in a jail setting. Rauschenberger email address: margaret.rauschenberger@alverno.edu.

Table 2.1 Elements for the performance assessment

Business and management program outcomes	Business and management external assessment outcomes
Uses discipline models and theories to analyze and communicate interdependence among systems, organizations, individuals and events	• Analyzes a business as a system of interdependent processes and makes strategic and operating decisions that optimize the system rather than its individual parts • Uses understanding of small business environment to assess the impact of a new business on the community and the socio/economic system • Effectively communicates business plan to meet the needs of a professional business audience
Applies business and management principles to develop and deliver quality products or services	• Conducts sound business research to evaluate a business idea • Integrates and applies principles and knowledge developed in previous courses, and designs a process to obtain relevant information to assess the feasibility of new product and service ideas
Uses team and organizational skills to work effectively with diverse individuals, teams and organizational units to meet stakeholder and organizational needs	• Integrates organizational and stakeholder perspectives to build a vision, objectives, and performance measures for a small business that reflects the needs of a variety of stakeholders • Balances personal and business goals with broader stakeholder needs to design a socially responsible business

Procedures. The assessment interview lasts about 20 min. The student then immediately completes a self assessment that highlights her mastery of the abilities and outcomes and asks her to do some self-reflection on her readiness for professional nursing practice. Once the student's self-reflection and self assessment is complete, the assessor provides some brief verbal feedback to the student. He or she then writes more extensive feedback that is provided to the student in an electronic format later the same day. Students who are unsuccessful on the performance assessment have some time for further preparation and allowed to reassess. The assessment is different because the assessor is not the same person.

Learning outcomes. This capstone assessment provides students with the opportunity to articulate accomplishments, using the language of the profession, to a member of that profession, and to demonstrate understanding of a need for continued growth and lifelong learning in nursing. A learning outcome is that the nursing candidate gains confidence in communicating her expertise to others.

Faculty learning outcomes. Faculty members study patterns in student performance over time from the professional nursing portfolio and the capstone performance assessment, as well as other student performance assessments. Each of these assessments is independently used to inform curriculum review and revision.

2.7.3 Alverno College School of Business Performance Assessment: Professional Interview and Business Plan[2]

The *Alverno College School of Business Performance Assessment: Professional Interview and Business Plan* is an externally administered performance assessment completed concurrently with the course, Small Business Management (MGT 400). It is administered by the Alverno College School of Business (AC 414). For the assessment, each student presents a business plan, which she/he has individually developed, to a local banker who assesses her professionalism, interpersonal communication, and general business knowledge. The interview is scheduled to follow the last session of the student's small business course. The candidate also completes a self-reflection and self assessment of her business plan, in addition to the results of the interview.

Synopsis of how students experience the assessment. This assessment is required for all undergraduate business majors in the Alverno College. Students complete an upper level, required small business course, and the *Alverno College School of Business Performance Assessment: Professional Interview and Business Plan* in their senior year. Midway through the small business course, the faculty coordinator for this performance assessment attends the small business class to describe the assessment, to create interest in and anticipation of the assessment, and to ask students to provide their bank location preference for their interview within the greater metropolitan Milwaukee and surrounding area. The School of Business makes every effort to match their location preference with a bank assessor in that geographic area, but does not promise to match the student with a specific banker or bank and does not permit the student to meet with a banker he or she already knows. At this point in the process, students realize that the interview will require them to stand alone and discuss their business plan and answer questions from the banker. This opportunity has tended to be a motivator for the student.

Purposes of the performance assessment. Purposes of the performance assessment are that faculty members receive an independent judgment from a trained assessor, a banker. Faculty members and also the banker, assess student integration and transfer of learning outcomes (integrated subject matter and learned abilities) for the business profession. During the semester, students have been actively engaged in independent research and development of a business plan for a new, small business venture. Thus, the purpose of the assessment is to provide students with an opportunity to demonstrate the advanced learning outcomes for their major field by using their analytic, problem solving, and social interaction abilities as integrated with the professional outcomes learned in the small business course, and to transfer them to an unfamiliar setting.

[2] Section prepared by Dena Lieberman, former Dean of the Alverno College School of Business and Professor of Business and Management. Lieberman earned her PhD in Anthropology from the University of Wisconsin, Madison, and an MBA from Marquette University. Lieberman email address: dena.lieberman@alverno.edu.

Learning outcomes. Learning outcomes for business undergraduates include their ability to identify opportunities for developing new and quality products and services to meet the changing needs of organizations and individuals. This learning outcome requires learners to use advanced level analytic and problem solving abilities developed in the business major and specifically in the small business and management course, where they develop a small business plan. The professional business interview and business plan requires the student to integrate and transfer her knowledge and abilities from a course setting to a new and unfamiliar professional setting, by creating an appropriate stimulus to call forth her integrated and transferable knowledge and abilities related to developing new products and services.

Further procedures. Toward the end of the semester, the coordinator for the assessment returns to the class to give students their banker assignment, to provide them some tips on setting up the bank interview, and to prepare for the interview regarding their business plan. At this point, students are often feeling more prepared with their business plan, because it has been almost completed. They are often both nervous and excited about the upcoming experience. The faculty coordinator assures them that prior students have found this to be a positive and productive learning experience, a criterion for Alverno performance assessments.

The performance assessment process and product. The product, the business plan, provides the foundation for a professional interview with a small business banker. In the interview, students have an opportunity to establish a banking relationship with the banker and to request financing for their venture. The banker in turn raises questions he or she would ask a new business loan applicant. These include: assessing potential risks, credit worthiness, and confidence in the business entrepreneur to carry out the business plan. The banker provides expert feedback about the business plan, the business idea, and corroborates feedback with supporting evidence. The banker also provides feedback on the student's business sense and overall professional conduct.

Student self-reflection and self assessment. The learner completes a self-reflection and self assessment of her business plan and of the results of the interview. Students comment on what they learned from the bankers, what they did well, and specific areas they could change and improve. Some students have refined their business plans based on the banker's feedback and have implemented their plans. The School of Business has also invited some students to enter business plan competitions and these students have won recognitions.

Elements for the performance assessment. Elements are derived from the learning outcomes for the School of Business major. The student is judged on the following:

- Dimensions of content knowledge integrated with learned abilities,
- Business sense,
- Professional demeanor,
- Focus on the business concept and innovative thinking,
- Knowledge of the market and competition,

Table 2.2 Business and management program outcomes related to external assessment outcomes

1. **Self-reflection on learning own abilities**. A faculty-stimulated learning process of using self assessment to engage learners in observing their performance of abilities, rather than their person. Learners develop expertise in their learning process as a learning outcome
2. **Self assessing performance and using feedback to transform it over time**. A learning process of deepening ones learning by interpreting, analyzing, and judging ones performance, using instructor and peer feedback to improve performance, gradually transforming it over time with diverse feedback from trained assessors (faculty, peers, or professional and business community partners). This learning process leads to self-confidence and self-efficacy as learning outcomes
3. **Learners developing metacognitive performance**. A learning process of recognizing patterns in knowledge-rich environments, thinking while performing and thinking about disciplinary frameworks, this transformative learning cycle assists learners to engage in restructuring their knowledge. Thus, learners are adapting and transferring disciplinary frameworks integrated with learned abilities to unscripted, unfamiliar settings in relation to diverse criteria and standards. In this transformative learning cycle, learners are thinking metacognitively, using their learned abilities to prompt Metacognitive Strategies as Frameworks for Performance. Thus, they engage their capacity for integrating *Reasoning* and *Performance* that are domains of the person. A transformative learning cycle that leads to learning outcomes in the disciplines and professions, learners are stimulated by a demonstrated cause of growth, that is, breadth of learning in Alverno's ability-based curriculum
4. **Learners developing professional expertise**. A learning process of engaging in reflective learning, envisioning improved role performance, and then monitoring role performance in relation to criteria and standards across multiple, varied situations to reach beyond what they thought they could do. Learners combine this element with using feedback from diverse sources to improve their performance over time. Learners are gradually experiencing this learning process as a transformative learning cycle, namely, Self Assessing Role Performance. In this way, learners are also drawing on their capacity for integrating *Performance* and *Self-Reflection* as domains of growth, building toward an identity as a learner, contributor, and professional, an outcome of Alverno's ability-based curriculum
5. **Learners developing identity as a self-sustained and unique learner, contributor, and professional**. A learning process of engaging in learning independently, somewhat before they engage and accomplish a developmental restructuring of their thinking and reasoning. This restructuring is leading learners toward interdependent learning, appreciating multiple perspectives, and mutuality. This learning process continues toward further developing mutuality, key in understanding one's own uniqueness in relationships with others, as one continues learning professionally. This learning process is gradual and developmental, and it seems that faculty members are essentially stimulating students to make these connections in a transformative learning cycle, namely, Engaging Diverse Approaches, Views, and Activities. This cycle is stimulated by learners' breadth of learning in Alverno's ability-based curriculum as a cause of growth, assisting learners to integrate *Self-Reflection* and *Development*, domains of growth

- Cash flow and ability to service debt,
- Business sense in responding to unanticipated questions, and
- Professional demeanor during the interview (see Table 2.2 on next page).

Student feedback on the performance assessment. Based on their feedback after the interviews, almost every student says the experience gave them confidence. This was because they were able to discuss their business plan in a professional setting and to learn more about what a banker is expecting in a business plan.

Training for quality of the assessor role. To ensure that bankers are able to perform their role in the interview and performance assessment, School of Business faculty members provide the banker with training on the purpose of the performance assessment. Faculty members also provide training with a set of criteria and examples, so that the bankers can provide quality, written feedback for the student.

Refinement of performance assessment over time. Over the years, the business faculty at Alverno College has developed a cadre of bankers who serve as trained assessors for this performance assessment interview and business plan. The process of maintaining an active group of trained bankers as assessors requires significant faculty resources because our local bank industry has undergone contractions and bankers are mobile.

Creating time for bankers to conduct a performance assessment. Faculty members estimate that a banker spends about four hours assessing one student: (1) reading the plan, (2) meeting with the student, and (3) providing written feedback. As expected, some bankers are more willing than others to devote the time for training and assistance with the interview that is required for a performance assessment.

Learning outcomes for assessors. The bankers who stay with the performance assessment process and product often comment that they view their work as part of their community service. They often note that they enjoy the experience of assisting students to understand what bankers expect in interviews about business plans. Bankers typically comment that they are impressed with the quality and comprehensiveness of the student interview and business plans.

Use of faculty time and administrative resources. When faculty members created the performance assessment years ago, they did not necessarily anticipate the effort required to maintain an engaged small business bank assessor pool. This performance assessment also required administrative support for follow-up with bankers to ensure that faculty members received their electronic feedback, and for uploading the feedback to each student's diagnostic digital student portfolio.

Thus, Alverno College School of Business faculty has learned that a purposeful effort is required to maintain a group of active bank assessors. Faculty members who accept the assignment to coordinate this assessment make personal calls to bankers at their workplace, maintain phone and email contacts to cultivate these relationships, and recruit and train new bank assessors as needed. Currently, the School of Business has over 30 active bank assessors in the assessor pool.

Use of student feedback and faculty judgment to select assessors. Faculty members have also invited students to provide feedback on the quality of their assessors' performance and to recommend whether the faculty should continue to ask a banker to serve as an assessor. Thus, student feedback, combined with faculty member judgment, has assisted the School of Business faculty to invite or pass on inviting a particular banker to assess. This strategy has assisted in maintaining a quality pool of trained assessors.

School of Business faculty perspective. The Alverno College School of Business faculty as a whole believes that the effort required for maintaining this performance assessment including a professional interview and business plan is worth continuing because the assessment challenges their students to integrate and transfer integrative knowledge and abilities learned in small business and management in order to demonstrate learning outcomes for the profession. Learning outcomes include: developing new and innovative products and services, researching business opportunities, analyzing relevant business information, and effectively interacting in professional business settings.

2.8 Alverno College School of Education Implementation of a Profession-Designed National Performance Assessment of Student Teaching

Three institutions have been participating in the profession-designed, nationally piloted, and field-tested implementation of the *edTPA* in Wisconsin (http://www.edtpa.com). The assessment proposes to create an inter-rater reliable and interstate instrument for establishing readiness for the profession of teaching by evaluating candidates teaching performance. The *edTPA* was jointly created by professionals at Stanford University, the Council of Chief State School Officers (CCSSO), and the American Association of Colleges of Teacher Education (AACTE). This performance assessment centers on candidates' self-reflection and documentation of a "learning segment" of linked learning experiences for their own students, and is completed during candidates' student teaching experiences.

Learning outcomes. The *edTPA* centers on the candidate's ability to (1) design meaningful plans for instruction, (2) engage students in instruction (measured in video recordings of lesson excerpts), (3) make sense of assessment and student learning data, and (4) reflect on what he or she has learned about the practice of teaching across the segment. While 80 % of the *edTPA* is identical across grade levels and content areas, 20 % of the assessment focuses on disciplinary and/or developmentally specific considerations (e.g., Performing Arts, Elementary Mathematics, Secondary English Language Arts, Special Education, and Early Childhood Education). The designers of the *edTPA* have selected Pearson to create and conduct the online evaluation system, but benchmarked and calibrated scorers will assess the candidates' work with professional teaching expertise in the content disciplines.

Purposes of the assessment. Alverno College School of Education chose the *edTPA* for implementation because (1) its conceptual terrain aligned with the faculty-designed performance assessment that candidates have consistently completed within the student teaching semester, namely *Teacher Effectiveness on*

Student Learning (TESL). This faculty-designed assessment has been validated internally (Rickards and Diez 1992) and externally (Zeichner 1997, 2000). The School of Education also advanced the implementation of the *edTPA* because (2) this performance assessment will be consequential for statewide, initial teacher licensure and continuous program review in September 2015.

Sample. Beginning in fall 2010, the entire student teaching cohort in general education, approximately 25–40 students per semester, completed the *edTPA* as part of their student teaching requirements. Starting in spring 2013, the entire student teaching cohort in special education also began completing the *edTPA*.

Wisconsin state policy. Wisconsin state policy requires 18 weeks of student teaching. Within those 18 weeks, student teacher candidates regularly complete two 9-week student teaching placements in each level of their licensure (e.g., developmental levels include kindergarten and elementary, middle, and high school). Alverno School of Education faculty members who supervise student teaching have implemented the *edTPA* during the first 9-week placement so that in cases where candidates may need to improve their performance, the improvements may be completed prior to Alverno's undergraduate commencement ceremony. (While student teacher placements follow a school's calendar, this event usually occurs three weeks prior to the completion of student teaching requirements.)

Alverno faculty used findings for improvement. Alverno School of Education faculty members have consistently used data for improvement (Diez 1988/1990; Merry et al. 2013). Thus, they used the *edTPA* data to evaluate program outcomes and make recommendations for curriculum revision. For example, supervisory faculty observed a weakness in candidates' asking of essential questions in elementary placements within middle childhood/early adolescent mathematics. Several candidates appeared to teach mathematics lessons less well than other candidates. Faculty members also noted that these candidates had not had an opportunity to teach a mathematics lesson prior to student teaching, because they had completed content areas outside of mathematics. Faculty addressed this issue by revising the course expectations in the field practicum immediately prior to student teaching, which now includes a "mini-*edTPA*", requiring candidates to document their teaching around their content certification areas over a 3-day period. School of Education faculty members have also used the *edTPA* data to examine other program outcomes, seeking to identify gaps and strengths among developmental levels and related content areas.

2.9 Validating the *edTPA*

Following completion of the *edTPA*, Alverno candidates reported that although the performance assessment was intensive, it accurately reflected their strengths as practitioners and required that they closely attend to their own students' or pupils' content learning outcomes. Candidates also reported that the depth and breadth of the commentaries for each *edTPA* component (e.g., Planning; Engaging Students and Supporting Learning; Assessing; Reflecting) did prepare them for the professional interview process, a main focus of the *edTPA*.

Alignment with areas of hiring. This profession-designed national performance assessment *aligned* with areas of hiring concerns in local school districts. Concerns included: (1) differentiating instruction from assessment for individuals and subgroups; (2) identifying essential understandings in instruction; (3) engaging students actively in learning experiences; (4) closely interrogating assessment data; and (5) connecting theory and practice.

Adaptation by Wisconsin institutions and their reports. Three Wisconsin institutions at the state level adapted the profession-designed national performance assessment. This experience was the subject of a recent statewide conference for the University of Wisconsin System. Pointer Mace (2012, September) reported on Alverno's experience with the *edTPA*. Pointer Mace acknowledged that since she was leading the statewide effort, her own preparation in the nuances of the *edTPA* may have been a factor in Alverno's relatively less problematic adaptation. (Marvin Lynn of University of Wisconsin, Eau Claire and Cheryl Hanley-Maxwell of University of Wisconsin, Madison also reported on their schools' implementation of the *edTPA.)*

Report on Alverno faculty experience. Pointer Mace noted that Alverno Education faculty attributed their relative ease of adaptation of the *edTPA* to the smaller number of students at the College, compared with the University of Wisconsin, Eau Claire and the University of Wisconsin, Madison. However, Alverno faculty also commented that the institutional culture of performance assessment may have been a factor affecting their experience. Further, the Alverno School of Education requires that each candidate complete a video analysis of a teaching sample, beginning in the first field practicum and continuing through student teaching. Since video analysis is a part of the *edTPA* experience, Alverno candidates may have experienced easier completion of the *edTPA*. Candidates also have experienced rubrics for self assessment, another possible factor that may account for faculty experience and learner use in implementing the *edTPA*.

2.10 Analysis of *edTPA* for Elements of Performance Assessment[3]

Pointer Mace analyzed the *edTPA* for the five elements of performance assessment. She used both the experiences of the Alverno faculty and post hoc analyses from student comments. She found the following elements:

[3] Section prepared by Desiree Pointer Mace, Associate Professor and Associate Dean, School of Education, Alverno College. She serves as a national design team member of the *edTPA*. Pointer Mace is also a member of the award-winning Valuing in Decision-Making Department. Pointer Mace earned her BA in Cognitive Science from Vassar College and PhD in Education from the University of California at Berkeley. Pointer Mace email address: desiree.Pointer-Mace@alverno.edu @dpointermace.

1. The first element, (1) *self-reflection,* was directly stimulated by the prompts embedded within the *edTPA* assessment handbooks. A candidate was required to analyze ones teaching in relation to the outcomes for student learning. As noted, candidates (1) design meaningful plans for instruction, (2) engage students in instruction (measured in video recordings of lesson excerpts), (3) make sense of assessment and student learning data, and (4) reflect on what he or she has learned about the practice of teaching across the segment.
2. The second element, (2) *self assessment*, occurred when candidates were invited to analyze their video samples of their teaching performances. For example, candidates completing the *edTPA* must identify points of connection (and lack thereof) regarding their own students' engagement within the video sample. Candidates must also analyze the effectiveness of their own performance assessment instruments, for capturing a continuum of their own students' understandings related to the content objectives. Candidates then created reasonable and warranted changes to their own learning segment design, as if they were to teach that particular content again.
3. The third element is (3) *learners developing professional expertise related to metacognitive performance.* Pointer Mace found that integrating Reasoning and Performance was evidenced in the *edTPA* performance data, not only through the content of the assessment but also by the process candidates used to create their documents. Candidates were able to create a timeline for completion of all required elements and hold one accountable. This is a critical professional capability for the work of classroom teaching.

   This finding was reinforced by a graduate of the program in a Major Forum on the *edTPA* conducted at the American Association of Colleges of Teacher Education (Whittaker et al. 2013). The graduate, Kathryn Miszewski, described the strong correlation between the performances evaluated by the *edTPA* assessment and the competencies critical to success in her first year of teaching, thus describing a relationship between the domains of her person, Reasoning and Performance. Thus, the *edTPA* more broadly addressed how the candidate conceptualized the role of the professional teacher, and how the candidate had developed metacognitive performance that connected the domains of her person, Reasoning and Performance.
4. The fourth element, (4) *learners developing professional expertise related to self assessing role performance,* was evidenced by the relationship of the *edTPA* components to the School of Education faculty-designed *Performance Assessment and Self assessment: Student Teaching Exit Portfolio.* This assessment encompasses the entire 18-week student teaching semester. It requires candidates to cumulatively self-reflect on and self assess their development as a teacher. The prompt for self assessment asks the candidate to articulate what she or he has learned as a teacher that is aligned with multiple disciplinary frameworks and abilities. The candidate is required to include the Wisconsin Standards for Educator Development and Licensure and (if a Special Education Candidate) the Council for Exceptional Children Standards. The student teacher is also required to use the Alverno Advanced Education Abilities as learning outcomes.

Pointer Mace's analysis demonstrated that one summative assessment, the *edTPA*, does not suffice for the fourth element, *developing professional expertise related to self assessing role performance*, nor should the *edTPA* be required to do so. The *edTPA* is also probably not valid for assessing the fifth element (5) *learners developing an identity as a self-sustained and unique learner, contributor and professional.* Both elements require continuous learning processes that are stimulated by faculty and cultivated across the entirety of the Alverno College Teacher Preparation Program and its performance assessments.

For example, at the end of the first undergraduate field practicum course, candidates engage in a performance assessment titled *I Have What It Takes.* This assessment requires them to closely read the dispositions of the Wisconsin Standards for Educator Development and Licensure, and provide evidence for their enactment, their performance, of selected dispositions. This orientation toward establishing that candidates successfully complete the assessment, *I Have What It Takes*, is not only elicited by but maintained throughout the program by faculty stimulation of students developing identity as a professional.

The *Alverno College School of Education Performance Assessment and Self Assessment: Student Teaching Exit Portfolio* stimulated candidates' *developing an identity as a self-sustained and unique learner, contributor, and professional.* Candidates provided evidence for their readiness to enter the profession with an identity as independent practitioners with high standards for their own learning as well as their own students' learning and development.

2.11 Scholarly Significance

Identifying effective elements of performance assessment based on conceptual frameworks in the external literature and Alverno literature, and derived from faculty practice, may lead to profession-designed and faculty-designed performance assessments that combine these elements (White et al. 2009). Professional identity formation may sustain effective performance over a lifetime of practice (Bebeau and Monson 2012; McKee and Eraut 2012).

So far, elements derived from a review of the external literature and Alverno literature included the following elements (see Table 2.2 for learning processes, transformative learning cycles, and learning outcomes). Faculty practice in performance assessment in the Schools of Nursing and Business and Management included the following five elements, and their attending learning processes. This is not a surprise, because they were designed to do so.

1. Self-reflection on learning own abilities.
2. Self assessing performance and using feedback to transform it over time.
3. Learners developing metacognitive performance.
4. Learners developing professional expertise.
5. Learners developing identity as a self-sustained and unique learner, contributor, and professional.

The profession-designed national performance assessment, *edTPA*, assessed the first three elements, according to faculty experience and candidate results. However, the School of Education had already created and validated a faculty-designed assessment. This final assessment for the student teaching experience, namely *Performance Assessment and Self Assessment: Student Teaching Exit Portfolio* captured element (4) *learners developing professional expertise*, with some support from the *edTPA*. Element (5) *learners developing identity as a self-sustained and unique learner, contributor, and professional,* is stimulated by additional performance assessments that learners complete throughout the School of Education professional program. These capture student development of both elements (4) and (5).

2.12 Conclusions

Faculty learning and student learning appear to occur for each of the five elements:

1. Self-reflection on learning own abilities.
2. Self assessing performance and using feedback to transform it over time.
3. Learners developing metacognitive performance.
4. Learners developing professional expertise.
5. Learners developing identity as a self-sustained and unique learner, contributor, and professional.

Learning processes occur for faculty and students with continuous performance assessment throughout their programs in nursing and business and management (see Table 2.3 and Fig. 2.1., Relationships Among Faculty Designs, Integrative Learning, Transfer of Learning, and Identity Development Prompted by Elements of Performance Assessment). These elements are made up of gradual learning processes and transformative learning cycles, and faculty members are stimulating students to make these connections so they develop learning outcomes. Faculty members appear to be learning from performance assessments (Mentkowski et al. 2012). Students and alumnae are learning from them, as evidenced by qualitative and statistical connections to the ability-based curriculum and its cultural context (Mentkowski et al. 2000; Rogers and Mentkowski 2004; Rogers et al. 2006).

Faculty-designed performance assessments illustrated all five elements, because they were designed to do so. The analysis of the *edTPA* found that each of the first three elements was experienced in Alverno faculty experience and by teacher candidates who commented on the assessment post hoc. However, faculty and candidates judged that element (4) was only related to the profession-designed national performance assessment when combined with the Alverno faculty-designed performance assessment, *Alverno College School of Education Performance Assessment and Self Assessment: Student Teaching Exit Portfolio*. Further, it appears that candidates had learned to use each of the five elements.

However, graduate professional schools should take note that Alverno learners were undergraduates and 5-year alumnae. Further, each professional school may

Table 2.3 Alverno College Bachelor of Science in Nursing Program Outcomes Related to Program Outcomes of the Alverno College School of Nursing Capstone Performance Assessment

BSN Program outcomes	BSN capstone assessment outcomes demonstrated by learner
Communicates creatively and effectively (Alverno ability: communication)	Communicates in multiple modes, using theories, strategies, and technology in professional practice
Integrates analytic frameworks within the practice of professional nursing (Alverno ability: analysis)	Identifies evidence of effective use of frameworks to address problems and meet the needs of clients
Applies problem solving processes to promote wellness in multiple environments (Alverno ability: problem solving)	Identifies evidence of her effective use of frameworks to address problems and meet the needs of clients
Uses valuing frameworks and ethical codes to promote human dignity (Alverno ability: valuing in decision-making)	Demonstrates the incorporation of values and ethics in decision-making
Interacts effectively in interpersonal, therapeutic, and group contexts (Alverno ability: social interaction)	Competently communicates the effect of individual qualities, qualifications, and environment on the success of therapeutic relationships
Advocates for and improves access to health care (Alverno ability: effective citizenship)	Identifies evidence of effective use of frameworks to address problems and meet the needs of clients
Fulfills the responsibilities of a professional practitioner in contemporary society (Alverno ability: developing a global perspective)	Consistently demonstrates characteristics of a professional nurse
Appreciates the uniqueness of self and others to promote wellness (Alverno ability: aesthetic engagement)	Demonstrates the incorporation of values and ethics in decision-making

choose to conduct their own analyses of faculty-designed and profession-designed performance assessments. The authors' have found that elements for performance assessment that rely on faculty and student learning are so broad as to be somewhat useful, which makes them more likely to be used. However, these elements may be *adapted* for use, because another curriculum, including Alverno's, *cannot be adopted*. We might all agree that the needs of a faculty and its student body are too unique but that rather, we might adapt from the external literature those ideas that prompt student learning that we may use with each institution's own students (Mentkowski 2006). It may be tempting to overgeneralize from the Alverno student and alumna learning outcomes because Alverno faculty and students learned from the ability-based curriculum and its performance assessments. Rather, we have set forth several elements for faculty and student learning that we believe are warranted by literature external to the College, Alverno literature, and our research findings.

Issue/Questions for Reflection

- As authors, we suggest that faculty teams might review a *profession-designed* or *faculty-designed* performance assessment for the elements set forth in this chapter
- Do our performance assessments include each of these elements? Are there some elements that we include that these authors do not?
- What learning processes do we stimulate for our learners with our performance assessments?
- Are there transformative learning cycles that we stimulate for our learners that these authors did not identify in the experiences of the Alverno faculty and their research findings? What are these, and how might we describe them for our students and ourselves?

Notes

Note 1. Institutional Review Board review is not applicable, because performance assessment is a component of the Alverno College program in each analysis of practice for nursing and business and management.

Note 2. Pointer Mace submitted the appropriate documentation to the Alverno Institutional Review Board for the *edTPA* study, and was approved by the Board. *Note 3*. Alverno students and alumnae contributed to a longitudinal study data collection from 1976–1990. These learning processes, transformative learning cycles, and learning outcomes are evidenced by qualitative and statistical connections to the ability-based curriculum and its cultural context (Alverno College Faculty 2000). Quantitative and qualitative data analyses were reported in Mentkowski et al. (2000). See Fig. 4.2. Alverno Curriculum as a Cause of Student Growth During College, pp. 128–129 and Fig. 4.3., Alverno Curriculum as a Cause of Sustained Learning and Growth After College, pp. 136–137. See also Appendix H: Developing Perspectives of Students on Self assessment, Using Feedback, Commitment to Improvement, and Role of Criteria, pp. 447–554. See also Appendix J: Five-Year Alumna Perspectives on Learning Outcomes and Causal Attributions to Alverno Curriculum and College Culture, pp. 455–457. Rogers and Mentkowski (2004) and Rogers et al. (2006) also confirmed these analyses and extended them.

References

Abromeit, J. M. (2012, April). Evaluating general education: Using a “home grown” student performance assessment. Paper for the Proceedings of the Annual Conference of the Higher Learning Commission, Chicago, IL.

Alverno College Council for Student Assessment. (2013). *Integration and transfer definitions and criteria from literature external to the College, Alverno literature, and Alverno faculty practice for faculty and student learning*. Milwaukee, WI: Alverno College Institute.

Alverno College Faculty. (1973/2010). *Ability-based learning program* [brochure]. Milwaukee, WI: Alverno College Institute (Original work published 1973, revised 1980, 1983, 1985, 1988, 1991, 1992, 1993, 1994, 1996, 2000, 2002, 2005, 2010).

Alverno College Faculty. (1976/2005). *Ability-based learning outcomes: Teaching and assessment at Alverno College*. Milwaukee, WI: Alverno College Institute (Original work published 1976, revised 1981, 1985, 1989, 1992, and 2005) (Before 2005, was *Liberal Learning at Alverno College*).

Alverno College Faculty. (1979/1994). *Student assessment-as-learning at Alverno College*. Milwaukee, WI: Alverno College Institute (Original work published 1979, revised 1985 and 1994).

Alverno College Faculty. (2000). *Self assessment at Alverno College*. In G. Loacker (Ed.). Milwaukee, WI: Alverno College Institute.

Alverno College Faculty. (2015). *Feedback IS teaching: Quality feedback for student and faculty learning*. Milwaukee, WI: Alverno College Institute.

Alverno College School of Education Performance Assessment and Self Assessment: Student Teaching Exit Portfolio. Milwaukee, WI: Alverno College Institute.

American Association of Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' views*. Washington, DC: Author.

American Nurses Association. (2001). *Code of ethics for nurses with interpretive statements*. Washington, DC: ANA.

American Nurses Association. (2003). *Nursing's scope and standards of practice*. Silver Spring, MD: ANA.

Anastasi, A. (1980). Abilities and the measurement of achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade*. New Directions for Testing and Measurement (Vol. 5, pp. 1–10). San Francisco: Jossey-Bass.

Argyris, C., & Schön, D. (1974). *Theory in practice: Increasing professional effectiveness*. San Francisco: Jossey-Bass.

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Baxter Magolda, M., & King, P. M. (2007). Constructing conversations to assess meaning-making: Self authorship interviews. *Journal of College Student Development, 48* (5), 491–508.

Bebeau, M. J., & Monson, V. E. (2012). Professional identity and formation across the lifespan. In A. McKee & M. Eraut (Eds.), *Learning trajectories, innovation and identity for professional development: Innovation and change in professional education* (Vol. 7). Dortrecht: Springer. Springer Science + B.V.

Boud, D., Keogh, R., & Walker, D. (Eds.). (1985). *Reflection: Turning experience into learning*. London: Kogan Page.

Boud, D., & Walker, D. (1998). Promoting reflection in professional courses: The challenge of context. *Studies in Higher Education, 23*(2), 191–206.

Boyatzis, R. E. (1982). *The competent manager: A model for effective performance*. New York: Wiley.

Boyatzis, R. E., Cowen, S. S., Kolb, D. S., et al. (1995). *Innovation in professional education: Steps on a journey from teaching to learning*. San Francisco: Jossey-Bass.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington, DC: National Academy Press.

Consortium for the Advancement of Teaching, Learning and Assessment: Shared Educational Assumptions. (1992, June). [Handout]. In Consortium for the Advancement of Teaching, Learning and Assessment. (1992, June). In Alverno College. (2012, June), *Connecting student learning outcomes to teaching, assessment, curriculum* (Section 12, p. 19). Milwaukee, WI: Alverno College Institute.

Diez, M. E. (1988/1990). A thrust from within: Reconceptualizing teacher education at Alverno College. *Peabody Journal of Education*, *65*(2), 4–18.

Eraut, M. (1994). *Developing professional knowledge and competence*. London: Falmer.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.

Ewell, P. T. (Ed.). (1985). *Assessing educational outcomes*. New Directions for Institutional Research (Vol. 47). San Francisco: Jossey-Bass.

Friedman, M., & Mentkowski, M. (1980, April). Validation of assessment techniques in an outcome-centered liberal arts curriculum: Empirical illustrations. Paper Presented at the Annual Meeting of the American Educational Research Association, Boston.

Friedman, M., Mentkowski, M., Deutsch, B., Shovar, M. N., & Allen, Z. (1982). Validating assessment techniques in an outcome–centered liberal arts curriculum: Social interaction generic instrument. Milwaukee, WI: Alverno College Productions (ERIC Document Reproduction Service No. ED 239 558).

Friedman, M., Mentkowski, M., Earley, M., Loacker, G., & Diez, M. (1980). Validating assessment techniques in an outcome-centered liberal arts curriculum: Valuing and communications generic instruments. Milwaukee, WI: Alverno College Productions (ERIC Document Reproduction Service No. ED 239 557).

Gruppen, L. D., White, C. J., Fitzgerald, T. J., Grum, C. M., & Wolliscroft, J. O. (2000). Medical students' self assessments and their allocations of learning time. *Academic Medicine, 75*(4), 374–379.

Hakel, M. D. & Halpern, D. F. (2005). How far can transfer go? Making transfer happen across physical, temporal, and conceptual space. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 357–370).

Hanley-Maxwell, C. (2012, September). Supporting special education candidates in the edTPA (Report from University of Wisconsin, Madison). University of Wisconsin System Conference on edTPA Implementation, Wisconsin Dells, WI.

Harward, D. W. (Ed.). (2011). *Transforming undergraduate education: Theory that compels and practices that succeed.* Lanham, MD: Rowman & Littlefield Publishers. http://www.rowman.com/ISBN/1442206748 (Ebook).

Kegan, R. (1982). *The evolving self: Problem and process in human development*. Cambridge, MA: Harvard University Press.

Loacker, G., Cromwell, L., & O'Brien, K. (1986). Assessment in higher education: To serve the learner. In C. Adelman (Ed.), *Assessment in American higher education: Issues and contexts* (Report No. OR 86-301, pp. 47–62). Washington, DC: U.S. Department of Education.

Loacker, G., Doherty, A., & Jensen, P. (2011/February). Competency/e in Business and at Alverno. [Handout]. In Alverno College. (2012, June), *Connecting student learning outcomes to teaching, assessment, curriculum* (Section 8, p. 13).

Loacker, G., & Jensen, P. (1988). The power of performance in developing problem solving and self assessment abilities. *Assessment and Evaluation in Higher Education, 13*(2), 128–150.

Loacker, G., & Mentkowski, M. (1982). A holistic view of adult abilities. In S. Ward & J. Kinney (Eds.), *Choices and issues in basic skills instruction* (Vol. 1, pp. 109–116). Washington, DC: National Institute of Education.

Loacker, G., & Palola, E. G. (Eds.). (1981). Clarifying learning outcomes in the liberal arts. *New Directions for Experiential Learning* (Vol. 12). San Francisco: Jossey-Bass.

Lynn, M. (2012, September). Supporting candidates in the *edTPA* (Report from University of Wisconsin, Eau Claire). In University of Wisconsin System Conference on *edTPA* Implementation, Wisconsin Dells, WI.

Marton, F., Hounsell, D., & Entwistle, N. (Eds.). (1984). *The experience of learning*. Edinburgh: Scottish Academic Press.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28*(1), 1–14.

McKee, A., & Eraut, M., (Eds.). (2012). *Learning trajectories, innovation and identity for professional development: Innovation and change in professional education* (Vol. 7). Dortrecht Heidelberg London New York: Springer. Springer Science + B.V. doi:10.1007/978-94-007-1724_1

Mentkowski, M. (1998). Higher education assessment and national goals for education: Issues, assumptions, and principles. In N. M. Lambert & B. L. McCombs (Eds.), *How students learn: Reforming schools through learner-centered education* (pp. 269–310). Washington, DC: American Psychological Association.

Mentkowski, M. (2006). Accessible and adaptable elements of Alverno student assessment-as-learning: Strategies and challenges for peer review. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 48–63). London, UK: Taylor and Francis.

Mentkowski, M., Abromeit, J., Mernitz, H., Talley, K., Knuteson, C., Rickards, W., et al. (2012, April). Assessing student learning outcomes across a curriculum: Resolving judgment and validity issues across disciplines and professions. In D. McKinley (Chair), *Clarifying assessment and measurement issues across disciplines and professions.* Symposium by Division I: Education in the Professions at the Annual Meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

Mentkowski, M., & Doherty, A. (1984). Abilities learned in college for later careering andlater success. In *American Association for Higher Education Bulletin*, (pp. 2 to 4 and 5 to 6).

Mentkowski, M., & Loacker, G. (1985). Assessing and validating the outcomes of college. In P. T. Ewell (Ed.), *Assessing educational outcomes.* New Directions for Institutional Research (Vol. 47, pp. 47–64). San Francisco: Jossey-Bass.

Mentkowski, M., & Sharkey, S. (2011). How we know it when we see it: Conceptualizing and assessing integrative and applied learning-in-use. In J. D. Penn (Ed.), *Assessing complex general education student learning outcomes.* New Directions for Institutional Research (Vol. no. 149). San Francisco: Jossey-Bass. doi:10.1002/ir.378 (wileyonlinelibrary.com)

Mentkowski, M., Moeser, M., & Strait, M. J. (1983). *Using the Perry scheme of intellectual and ethical development as a college outcomes measure: A process and criteria for judging student performance* (Vols. 1 and 2). Milwaukee, WI: Alverno College Productions.

Mentkowski, M., et al. (2000). *Learning that lasts: Integrating learning, development, and performance in college and beyond*. San Francisco: Jossey-Bass.

Merry, S., Price, M., Carless, D., & Taras, M. (Eds.). (2013). *Reconceptualizing feedback in higher 945 education: Developing dialogue with students*. Abingdon, Oxfordshire, UK: Routledge. 946

Mestre, J. P. (Ed.). (2005). *Transfer of learning from a modern multidisciplinary perspective.* Greenwich: CN: Information Age Publishing, Inc.

Perry., W. G, Jr. (1970). *Forms of intellectual and ethical development in the college years: A scheme*. New York: Holt, Rinehart & Winston.

Perry, W. G., Jr. (1980, January). Do students think we mean what we think we mean? Presentation at Alverno College Faculty Institute, Milwaukee, WI.

Pointer Mace, D. (2012, September). Assessment as opportunity: Using *edTPA* data for program review and improvement by Alverno College School of Education faculty. University of Wisconsin System Conference on *edTPA* Implementation, Wisconsin Dells, WI.

Read, J., & Sharkey, S. R. (1985). Alverno College: Toward a community of learning. In J. S. Green, A. Levine, et al. (Eds.), *Opportunity in adversity: How colleges can succeed in hard times* (pp. 195–214). San Francisco: Jossey-Bass.

Rickards, W. H., & Diez, M. E. (1992, April). Integrating multiple internal and external data sources in the institutional evaluation of teacher education. In M. Woodward (Chair), *Research and evaluation in graduate and professional education.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. Milwaukee, WI: Alverno College Productions.

Rogers, G., & Mentkowski, M. (2004). Abilities that distinguish the effectiveness of five-year alumna performance across work, family, and civic roles: A higher education validation. *Higher Education Research & Development, 23*(3), 347–374.

Rogers, G., Mentkowski, M., & Resetter Hart, J. (2006). Adult holistic development and multidimensional performance. In Carol Hoare (Ed.), *Handbook of Adult development and learning*. New York: Oxford University Press.

Schön, D. A. (1973). *Beyond the stable state. Public and private learning in a changing society*. Harmondsworth: Penguin.

Schön, D. (1983). *The reflective practitioner: How professionals think in action*. London: Temple Smith.

Schön, D. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass.

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational Researcher, 27* (3), 11–20.

White, C. B., Ross, P. T., & Gruppen, L. D. (2009). Remediating students' failed *OSCE* performances at one school: The effects of self assessment, reflection, and feedback. *Academic Medicine, 84*(5), 651–654.

Whittaker, A., Pointer Mace, D., Miszewski, K., Lys, D., Tomlinson, L., & Berry, B. (2013). Voices from the field: The impact of *edTPA* on new teachers and teacher preparation programs. Major Forum at the annual meeting of the American Association of Colleges of Teacher Education, Orlando, FL. Retrieved March 22, 2013 from http://vimeo.com/60864344

Zeichner, K. M. (1997, August). *Ability-based teacher education: Elementary teacher education at Alverno College.* New York: National Center for Restructuring Education, Schools, and Teaching (NCREST), Teachers College, Columbia University.

Zeichner, K. (2000). Ability-based teacher education: Elementary teacher education at Alverno College. In L. Darling-Hammond (Ed.), *Studies of excellence in teacher education: Preparation in the undergraduate years* (pp. 1–66). Washington, DC: American Association of Colleges for Teacher Education.

Chapter 3
Improving the Integration of Performance Assessments Within the Curriculum Across the Professions

Ilene Harris

Abstract Assessment of performance is an essential component of any curriculum across the professions. In this chapter, we characterize the major conceptual frameworks and traditions of development and investigation for the curriculum. In this context, we provide recommendations for policies and best practices for improving the integration of performance assessment within curricula across the professions. Other chapters in this book explain the importance of an array of different types of assessment. This chapter provides a broader context for consideration of the role of assessment generally, as integrated into curricula across the professions. A key recommendation is that assessment of performance in the professions includes both specialized and subspecialized knowledge for the profession including practical skills domains, and equally important, other cross professional competencies, such as professionalism, communication skills, collaboration and teamwork, and skills for reflective practice, and lifelong learning. In turn, another key recommendation is that for assessment in the context of professions curricula, assessment of performance in the actual settings of practice, or in authentic simulations, is essential.

Takeaways

- Assessment of performance in the professions should include assessment of specialized knowledge and skills, but also cross professional competencies, such as professionalism, communication skills, collaboration and teamwork, and skills for practitioner reflective practice and lifelong learning.
- Assessment of performance in professions curricula should include assessment in the actual settings of the workplace or in authentic simulations.
- Assessment of performance in the professions should be considered in the context of three complementary perspectives on curriculum: the 'systems' perspective which focuses on alignment of goals, instruction and

I. Harris (✉)
Faculty of Education, University of Illinois at Chicago, Chicago, IL, USA
e-mail: ibharris@uic.edu

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_3

assessment; the ‘reconceptualist’ perspective which focuses on assessment in the institutional and societal context of the curriculum and the ‘experienced’ and ‘hidden curriculum’; and the ‘deliberative inquiry’ perspective which focuses on stakeholder group reflective inquiry about appropriate assessment in relation to desired learning outcomes.

3.1 Conceptual Frameworks for Curriculum Design and Implementation

We first provide a broad and comprehensive conception of the term “curriculum”. “Curriculum” does not refer solely to the content or subject matter of an education program. Consistent with current conceptions of curriculum practice and scholarship, “curriculum” refers more broadly to every facet of the planning and implementation of education programs including: general and targeted needs assessment; formulation of learning goals and objectives; selection of approaches and methods of instruction including the teaching and learning environment; assessment of learners’ performance; and evaluation of the education program (Pinar et al. 1996; Schubert et al. 2002). Further, the term “curriculum” encompasses the concept of the “experienced curriculum”, what learners actually experience in education settings (Hafferty and Hafler 2011) and the “hidden curriculum”, the “unintended curriculum” and the “informal curriculum” (Pinar et al. Lave and Wenger 1991), all referring to the ‘experienced’ curriculum, which may or may not be consistent with the planned, formal curriculum and which may or may not be positive and effective for learning and professional practice (Harris 2011). For example, the planned and formal curriculum for professional education typically includes learning goals and objectives focused on communication skills, professionalism, and interprofessional teamwork. Yet, professionals, serving as role models in the setting of the workplace may exhibit poor communication skills, lack of professionalism, and poor teamwork (Castellani and Hafferty 2006; Goold and Stern 2006; Hafferty and Franks 1994; Hafferty 1998, 1999, 2000; Hundert et al. 1996; Stern and Papadakis 2006).

In this context, it is important to observe, from the perspectives of learning and assessment, that curricula for the professions typically include learning in the practice settings of the workplace, early in the formal curriculum, as early as the first year. Thereby, the formal curriculum confronts, powerfully, the “hidden curriculum”, the “informal curriculum”, and the “unintended curriculum”, as just indicated, evident in the role modeling and systems of professional practice in the setting of the workplace.

Curricula for the professions include early learning in the workplace for the purpose of achieving essential goals of education for the professions. For example, Shulman (2005) conceptualized professional development in terms of apprenticeships of the heart, i.e., professional socialization; of the hand, i.e., professional skills; and the mind, i.e., specialized knowledge, through what he referred to as *signature pedagogies*. In education for the professions, learning in the practice settings of the workplace is the signature pedagogy, where novices are socialized into a community of practice; develop professional skills through observation, role modeling, practice, reflection and feedback; and develop the motivation and context for application of knowledge learned in classroom settings (Harris 2011).

In turn, early learning in the workplace provides essential experiences that are consistent with developing the competencies needed for professional practice given the nature of professional practice (Schon 1987). Clearly, each profession requires specialized and often subspecialized knowledge. But in addition, professional practice situations are characterized by conditions of complexity, uniqueness, uncertainty, ambiguity, and conflicting value orientations (Harris 2011). In turn, effective professional practice requires 'practical knowledge' for applying a repertoire of specialized knowledge in various specific situations and related reflective competencies for self-assessment, responding to others' assessments, independent learning and self-correction, typically learned in practice, through experience with the actual problems of practice.

It is noteworthy that an evolving array of learning theories supports inclusion of early learning in the workplace as a component of the curriculum for education in the professions. Behavioral learning theories focus on the importance for professions education of opportunities for practice, with supervision, and specific and timely feedback in relation to observation of learners' performance in practice (Wilkerson and Irby 1998). Constructivist and 'adult' theories of learning focus on the importance for professional education of active learning by self-directed learners in relation to authentic problems; with opportunities to practice in multiple situations; to reflect upon their experiences and formulate principles for practice based on their reflections; and to apply these principles to practice in new situations (Kolb 1984; Schon 1987; Shuell 1986; Sternberg and Wagner 1986). Social learning theories also focus on the importance of active learning by self-directed learners. In addition, they focus on the importance of constructing meaning through observation and collaboration with others in the context of the workplace (Bandura 1986; Salomon and Perkins 1998; Vygotsky and Cole 1978). In this context, they learn in, and are initiated into, authentic communities of practice, beginning as novices, with legitimate peripheral participation and progressing to increasingly full participation in the community of practice (Brown et al. 1989; Collins et al. 1989; Lave and Wenger 1991; Wenger 1998).

We have provided a broad and comprehensive conception of the concept and practice of the "curriculum". We have observed that curricula for the professions typically include learning in the practice settings of the workplace, early in the formal curriculum, as early as the first year, to achieve fundamentally important goals of professional education, an approach to education that is consistent with an array of learning theories. In this context, it is clear that opportunities for authentic assessment are available and should be integrated in curricula for the professions. In the next section, we characterize the major conceptual frameworks and traditions of development and investigation for the curriculum. In this context, we will provide recommendations for policies and best practices for improving the integration of performance assessment within curricula across the professions. There are three predominant, and ultimately complementary, conceptual frameworks and traditions of development and investigation in curriculum studies: the ‘systems’ approach; ‘deliberative curriculum inquiry’; and ‘reconceptualist inquiry’ (Harris 1993a).

The ‘systems’ approach focuses on the systematic selection and alignment among various components of curricula, to include: general and targeted needs assessment; learning goals and objectives; instructional and learning methods; assessment methods; and evaluation methods. The ‘deliberative inquiry’ approach focuses on group process of reflective inquiry in curriculum development, processes often used to make justified decisions about the curriculum components in the ‘systems’ approach. The ‘reconceptualist’ approach focuses on use of perspectives and methods from a range of disciplines, such as ethnography, political science, and economics, to explore: (1) the relation between curricula and their contexts, such as their economic, political, social and cultural contexts and (2) the experienced and hidden curriculum.

3.2 Conceptual Frameworks for Curriculum Development and Scholarship: Best Practices for Improving the Integration of Performance Assessment in Curricula Across the Professions: The Systems Approach

The systems approach has provided professional education, and indeed education in general, with its most pervasive conceptual framework for curriculum development and evaluation. Synthesized by Tyler (1949) and in iconic texts in the various professions, such as Kern’s and colleagues’ text (Kern et al. 2009), the systems approach focuses on the: systematic analysis of needs; systematic formulation of purposes, goals and objectives; selection and organization of teaching and learning experiences; assessment of students’ performance; and evaluation of the curriculum. This approach has been “translated into a quasi-technology” (Harris 2011) grounded in the ideal of formulating specific behavioral objectives (Bloom 1956; Gronlund 1991; Krathwohl 1964).

In the ‘systems’ approach to curriculum design and implementation, fundamental issues relate to selection of appropriate learning experiences for helping students to achieve the learning goals and objectives and appropriate approaches for assessment of their achievement. For example, for development of specialized knowledge pertinent to each profession, formal education sessions, e.g., classroom learning, may be most appropriate, although even in the context of classroom learning, there has been widespread implementation of student-centered and case-based strategies for instruction, such as problem-based learning (Neville and

Norman 2007) and team-based learning (Thompson et al. 2007) consistent with constructivist learning theories. For assessment, there has been increasingly widespread use of written tests in constructed response formats, i.e., multiple choice questions, typically using case-based scenarios, to assess applied knowledge at higher cognitive levels with validity and reliability (Downing 2009; Case and Swanson 1998) and recommendations for best practices in constructing and analyzing the results from written tests (Downing 2009; Hawkins and Swanson 2008). For development of applied knowledge and procedural skills, simulation used for both instruction and assessment, may be the most appropriate type of assessment, for providing a safe environment with standardized problems for learning and assessment prior to learning in the practice settings of the workplace (Yudkowsky 2009; McGaghie and Issenberg 2009; Scalese and Issenberg 2008).

Ultimately, for education in the professions, learning and assessment in the context of the workplace are essential for development and application of specialized knowledge and procedural skills. Moreover, learning and assessment in the practice setting of the workplace serve multiple purposes in relation to professional learning goals and objectives: to develop professional competencies; to provide for initiation and socialization into a community of practice; to provide motivation for ‘classroom’ learning; and to provide an authentic context for understanding implications of learning for professional practice (Harris 2011). It has been argued that the learning goals and objectives best served by learning and assessment in the practice setting of the workplace are goals in the domain of professional development and socialization rather than cognitive knowledge goals and objectives (Harris 2011). Fundamentally, the workplace provides a venue where novices may observe “best practices” of master practitioners, reflect with master practitioners about what they have observed, engage in professional practice, themselves, at a level appropriate to their experience and training; and reflect on their own practice efforts with feedback from these master practitioners. In the practice settings of the workplace, assessment of learners’ practice, based on observation, is used for formative purposes for learning. In addition, assessment in the context of the workplace also provides a venue for more formal high stakes assessment and feedback, based on observation, and recorded with evaluation forms using global rating scales, checklists, and narratives (McGaghie et al. 2009; Pangaro and Holmboe 2008; Holmboe 2008).

Despite its clear logic and successful application in practice, the systems approach to curriculum design and implementation, particularly for education for the professions, has limitations, when used as the sole approach for curriculum development and scholarship. First, the systems approach does not clearly acknowledge or emphasize that professional practice learned ultimately in the workplace requires use of judgment and action in complex situations. In addition to conceptual and technical knowledge and skills, professionals need to develop and demonstrate reflective and practical competencies for dealing with problems in the indeterminate areas of practice that do not yield to technical or familiar solutions (Schon 1987; Harris 2011). Such competencies, which are difficult to describe in behavioral objectives need to be practiced, observed, and assessed in authentic simulations or in the actual settings of professional practice (McGaghie et al. 2009; Pangaro and Holmboe 2008; Holmboe 2008). Second, the systems approach does not explicitly acknowledge, or emphasize, that curricula are embedded in complex institutional settings, typically with multiple missions, including the provision of service to ‘clients’ and society, education, and often research and scholarship. Third, the systems approach to curriculum development tends to virtually ignore the nature of the informal and “hidden” curricula present in every institution of professional practice, most important, the role modeling of master practitioners of medicine, of law, of engineering, of divinity, and so on, which may not be consistent with “best practices”, recommended practices or even appropriate practices that are part of the canons of curriculum recommendations (Harris 2011).

3.3 Conceptual Frameworks for Curriculum Development and Scholarship: Best Practices for Improving the Integration of Performance Assessment in Curricula Across the Professions: The ‘Reconceptualist’ Approach

A more recent approach to curriculum development and scholarship that address these issues has been referred to as the ‘reconceptualist’ approach. This label is derived from the seminal work of a group of curriculum scholars who sought to ‘reconceptualize’ how we think about the curriculum (Shubert et al. 2002; Pinar et al. 1996). They recommended, and they and their followers used, conceptual frameworks and perspectives from a range of disciplines, such as ethnography, sociology, political science, and economics. They used these diverse conceptual frameworks and methods to focus on the relationships between curricula and the cultural, social, political, and economic structures of the professional school and workplace setting; the hidden curriculum of role modeling and professional socialization; and the curriculum that students actually experience (Harris 1993a, b, 2011).

In turn, a number of studies have explored these aspects of professional development. For example, investigators have studied and powerfully discussed the role of the hidden curriculum in development of health professionals in both academic institutional settings and the practice settings of the workplace (Castellani and Hafferty 2006; Goold and Stern 2006; Hafferty and Franks 1994; Hafferty 1998, 1999, 2000; Hundert et al. 1996; Stern and Papadakis 2006). These studies almost uniformly demonstrate that the professional values recommended in the formal curriculum, are not in fact, consistently demonstrated in the practice settings of the workplace. For example, Stern (1998) reports a study comparing the “recommended curriculum” of medical values, identified through content analysis of curriculum documents, with the values actually taught in hospital-based internal medicine teams at an academic medical center identified through naturalistic but systematic observation. Among his findings was that while the formal curriculum emphasized the importance of interprofessional respect, the naturalistic observation of actual practice in the setting of the workplace demonstrated pervasive professional disrespect.

The ‘reconceptualist’ approach to curriculum development and scholarship is clearly complementary to the ‘systems’ approach in focusing our attention on the institutional and societal context of the curriculum; and on the ‘experienced curriculum’ and the ‘hidden curriculum’. The implications for assessment relate to the professional competencies that must be assessed, principally the competencies to function in complex institutional settings and the practice settings of the workplace; and the competencies to engage in reflective practice in the evaluation of role models in the workplace. As discussed, such competencies which are difficult to describe in behavioral objectives need to be practiced, observed, and assessed in authentic simulations or in the actual settings of professional practice (McGaghie et al. 2009; Pangaro and Holmboe 2008; Holmboe 2008).

3.4 Conceptual Frameworks for Curriculum Development and Scholarship: Best Practices for Improving the Integration of Performance Assessment in Curricula Across the Professions: The Deliberative Inquiry Approach

Another recent, and also complementary, approach to curriculum development and scholarship that does address these issues has been referred to as ‘deliberative curriculum inquiry’, with recommendations for processes for curriculum deliberation and decision-making, fundamentally group processes of reflective inquiry (Harris 1991, 1993a, 2011). This approach has its origins in work by Schwab at the University of Chicago in the early 1970s (Schwab 1978). He viewed curriculum issues as essentially practical problems about making choices and taking action in complex situations, in local contexts, in which values and belief systems play a central role, as well as conceptual and applied knowledge. An example of such problems, more generally, is the issue of where to build an airport, given the possible disruption of communities. He argued, therefore, that curriculum problems should be addressed by methods appropriate to issues of choice and action, namely deliberation among stakeholders, such as education leaders, faculty members, students, and members of the community, who bring diverse perspectives and values about what to teach, why and how to teach and how to assess; and in turn reach a consensus based on their deliberative processes (Harris 1991, 1993a, 2011).

Schwab and others have characterized processes and “arts” (Harris 1991) for leading and participating in productive curriculum deliberations. The fundamental process is a systematic method by which properly constituted groups, stakeholders in a particular setting, formulate and consider alternative perspectives and formulations of education problems in a particular setting, as well as a variety of alternative solutions, about what and how to teach in a particular context.

Clearly, the most important deliberations for professional education take place at each professional school, involving key stakeholders, including: faculty members in the academic setting; representatives from the practice setting of the workplace, including workplace supervisors; administrators and education leaders; students and a chairperson. Heuristics have been developed to facilitate the group deliberative processes. Among the heuristics are preplanning evaluation, a process of preliminary data gathering to inform stakeholders in all phases of the deliberation (Curry 1992) and nominal group technique, a structured group process approach developed in management sciences (Hegarty 1971).

In addition, various perspectives should be used to inform the process of deliberation. Clearly, first and foremost, the group should consider the curriculum already in place and the local circumstances, for example: the missions, goals, and objectives of the institution; the characteristics of students; the orientations of faculty members; the resources available; the social, cultural, economic and political context; and the strengths and pressing problems—all based on the diverse perspectives of the group and informed by curriculum evaluation data (Bordage and Harris 2011). Second, the group should have knowledge of effective processes of curriculum design, such as the value of alignment among needs assessment; learning goals and objectives; and instructional methods and assessment methods, embedded in the systems approach to curriculum design (e.g., Tyler 1949; Kern et al. 2009). The learning objectives need not necessarily be expressed in terms of behavioral objectives; the methods of education extend to learning in the practice settings of the workplace; and assessment encompasses the very broadest goals for professional education (Harris 2011).

Third, the group should be knowledgeable about current thinking about the nature of professional practice, such as Schon's work on reflective (Schon 1983, 1987) and studies of the development of expertise and professional competence indicating the importance of learning and assessment in the authentic settings of the workplace (e.g., Bandura 1986; Brown et al. 1989; Lave and Wenger 1991). As another example, Ericsson's theory of expertise is useful to guide instructional development. According to Ericsson (Ericsson et al. 1993; Ericsson 2004), expertise develops through 'deliberative mixed practice with feedback', clearly the use of assessment for learning. In recent work, Mylopoulos and Regehr argue that for medical students to develop the skills for designing new solutions in complex workplace settings, we need to consider theories of adaptive expertise for innovative problem solving as a core competency (Myopoulos and Regehr 2009). Fourth, the group should consider perspectives in the national arena for their profession, such as the movements in medicine and nursing, related to professionalism, evidence-based health care, and patient safety and quality improvement. Fifth, the group should consider perspectives and evidence related to the "hidden curriculum", the experience of professional education in the practice settings of the workplace. As discussed previously, the group should compare their recommendations for the formal curriculum, such as applied knowledge of patient safety and quality improvement, with the actual culture of quality improvement in actual health care systems.

With regard to the purposes and functions of curriculum deliberation, Harris (2011) comments:

> These processes, if conducted well, serve multiple purposes. This process has the potential to bring together diverse values and sources of evidence and expertise; to reach justified decisions; to educate participants through exploration of diverse perspectives; and to achieve personal and political commitment to decisions. Clearly, these processes, by involving diverse stakeholders—including preceptors and mentors in workplace settings—have the potential to create curricula suitable for education in the setting of the workplace, whether the formal curriculum, the informal curriculum or the "hidden" curriculum of professional education. (p. 12)

In the 'deliberative curriculum inquiry' approach to curriculum design, the implications for assessment are that the group must be knowledgeable about appropriate approaches to assessment, particularly for the competencies they view as essential outcomes (McGaghie et al. 2009; Pangaro and Holmboe 2008; Holmboe 2008).

3.5 Conclusion

Assessment of performance is an essential component of any curriculum across the professions. In this chapter, we characterized the major conceptual frameworks and traditions of development and investigation for the curriculum. In this context, we provided recommendations for policies and best practices for improving the integration of performance assessment within curricula across the professions. A key recommendation is that assessment of performance in the professions includes both specialized and subspecialized knowledge for the profession including practical skills domains, and equally important, other cross professional competencies, such as professionalism, communication skills, collaboration and teamwork, and skills for reflective practice and lifelong learning. In turn, another key recommendation is that for assessment in the context of professions curricula, assessment of performance in the actual settings of practice, or in authentic simulations, is essential early during the curriculum, just as learning in the practice settings of the workplace is now common, even during the first year of the curriculum.

Issues/Questions for Reflection

- What resources are necessary, and potentially available, for implementation of recommended best practices for assessment across the professions?
- How important is it to assess cross professional competencies, such as professionalism, communication skills, and collaboration and teamwork?
- What are the most important issues in assessment of professional competence in the setting of the workplace, as recommended?

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bloom, B. S. (1956). *Taxonomy of educational objectives: Cognitive domains*. New York: McKay.

Bordage, G., & Harris, I. (2011). Making a difference in curriculum reform and decision-making processes. *Medical Education, 45*, 87–94.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32–42.

Case, S., & Swanson, D. (1998). *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners.

Castellani, B., & Hafferty, F. (2006). Professionalism and complexity science: A preliminary investigation. In D. Wear & J. M. Aultman (Eds.), *Medical professionalism: A critical review* (pp. 3–23). New York: Springer.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser (pps* (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Curry, L. (1992). Deliberative curriculum inquiry and its application in the education of health service administrators. *Journal of Health Administration Education, 10*, 519–526.

Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professional education* (pp. 149–184). New York: Routledge.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 82*, 370–374.

Goold, D. S., & Stern, D. T. (2006). Ethics and professionalism: What does a resident need to learn? *American Journal of Bioethics, 6*, 9–17.

Gronlund, N. E. (1991). *How to write and use instructional objectives*. New York: MacMillan.

Hafferty, F. W., & Franks, R. (1994). The hidden curriculum, ethics teaching, and the structure of medical education. *Academic Medicine, 69*, 681–871.

Hafferty, F. W. (1998). Beyond curriculum reform: Confronting medicine's hidden curriculum. *Academic Medicine, 73*, 403–407.

Hafferty, F. W. (1999). Managed medical education. *Academic Medicine, 74*, 972–979.

Hafferty, F. W. (2000). In search of a lost cord: Professionalism and medical education's hidden curriculum. In D. Wear & J. Bickel (Eds.), *Educating for professionalism: Creating a culture of humanism in medical education* (pp. 11–34). Iowa City: University of Iowa Press.

Hafferty, F. W., & Hafler, J. P. (2011). The hidden curriculum, structural disconnects, and the socialization of new professionals. In J. P. Hafler (Ed.), *Extraordinary learning in the workplace*. Dordrecht: Springer.

Harris, I. (1991). Deliberative inquiry: The arts of planning. In E. C. Short (Ed.), *Forms of curriculum inquiry* (pp. 287–321). Albany, NY: State University of New York Press.

Harris, I. (1993a). Perspectives for curriculum renewal in medical education. *Academic Medicine, 68*, 484–486.

Harris, I. (1993b). New expectations for professional competence: Reflective practice and self-reflection. In L. Curry & J. Wergin (Eds.), *Educating professionals* (pp. 17–51). San Francisco: Jossey-Bass.

Harris, I. (2011). Conceptual perspectives and the formal curriculum. In J. P. Hafler (Ed.), *Extraordinary learning in the workplace*. Dordrecht: Springer.

Hawkins, R. E., & Swanson, D. B. (2008). Using written examinations to assess medical knowledge and its application. In E. S. Holmboe & E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 42–59). Philadelphia, PA: Mosby - Elsevier.

Hegarty, E. (1971). The problem identification phase of curriculum deliberation: Use of the nominal group technique. *Journal of Curriculum Studies, 9*, 31–41.

Holmboe, E. S. (2008). Direct observation by faculty. In E. S. Holmboe & E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 119–129). Philadelphia, PA: Mosby - Elsevier.

Hundert, E. M., Hafferty, F., & Christakis, D. (1996). Characteristics of the informal curriculum and trainees' ethical choices. *Academic Medicine, 71*, 624–642.

Kern, D. R., Thomas, P. A., Howard, D. M., & Bass, E. B. (2009). *Curriculum development for medical education: A six-step approach*. Baltimore, MD: Johns Hopkins University Press.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Krathwohl, D. R. (1964). *Taxonomy of educational objectives: Affective domain*. New York: McKay.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

McGaghie, W. C., & Issenberg, S. B. (2009). Simulations in assessment. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professional education* (pp. 245–268). New York: Routledge.

McGaghie, W. C., Butter, J., & Kaye, M. (2009). Observational assessment. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professional education* (pp. 185–216). New York: Routledge.

Myopoulos, M., & Regehr, G. (2009). How student models of expertise and innovation impact the development of adaptive expertise in medicine. *Medical Education, 43*, 127–132.

Neville, A. J., & Norman, G. R. (2007). PBL in the undergraduate MD program at McMaster University: Three iterations in three decades. *Academic Medicine, 82*, 370–374.

Pangaro, L., & Holmboe, E. S. (2008). Evaluation forms and global rating scales. In E. S. Holmboe & E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 24–41). Philadelphia, PA: Mosby - Elsevier.

Pinar, W. F., Reynolds, W. M., Slattery, P., & Taubman, P. M. (1996). *Understanding curriculum*. New York, NY: Lang.

Salomon, G., & Perkins, D. N. (1998). Individual and social aspects of learning. In P. D. Pearson & A. Iran-Nehad (Eds.), *Review of Research in Education*. Washington DC: American Educational Research Association.

Scalese, R. J., & Issenberg, S. B. (2008). Simulation-based assessment. In E. S. Holmboe & E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 179–200). Philadelphia, PA: Mosby - Elsevier.

Schon, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schon, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey Bass.

Schwab, J. J. (1978). The practical: A language for curriculum. In I. Westbury & N. Wilkof (Eds.), *Science, curriculum and liberal education: Selected essays* (pp. 287–321). Chicago: University of Chicago Press.

Schubert, W. H., Schubert, A. L., Thomas, P., & Carroll, W. M. (2002). *Curriculum books: The first hundred years* (2nd ed.). New York, NY: Lang.

Shuell, T. J. (1986). Cognitive conceptions of learning. *Review of Educational Research, 56*, 411–436.

Shulman, L. S. (2005). Signature pedagogies in the professions. *Daedalus, 134*, 52–59.

Sternberg, R. J., & Wagner, R. K. (Eds.). (1986). *Practical intelligence: Nature and origins of competence in the everyday world*. New York: Cambridge University Press.

Stern, D. (1998). Practicing what we preach? An analysis of the curriculum of values in medical education. *The American Journal of Medicine, 104*, 569–575.

Stern, D. T., & Papadakis, M. (2006). The developing physicians—becoming a professional. *New England Journal of Medicine, 355*, 1794–1799.

Thompson, B. M., Schneider, V. F., Haidet, P., Levine, R. E., McMahon, K. K., Perkowski, L. C., et al. (2007). *Medical Education, 41*, 250–257.

Tyler, R. (1949). *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.

Vygotsky, L. S., & Cole, M. (Eds.). (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identify*. New York: Cambridge University Press.

Wilkerson, L., & Irby, D. (1998). Strategies for improving teaching practice: A comprehensive approach to faculty development. *Academic Medicine, 73*, 387–396.

Yudkowsky, R. (2009). Performance tests. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professional education* (pp. 217–244). New York: Routledge.

Chapter 4
Beyond Authenticity: What Should We Value in Assessment in Professional Education?

Christopher O'Neal

Abstract Authenticity assessments evaluate learners using methods and contexts that mimic the way the tested content and skills will be used in the real world. While authenticity has long been a goal of assessors across the education spectrum, educators have struggled with the supposed tradeoff inherent to authentic assessment: reliability versus validity. This tradeoff was particularly concerning in the large-scale assessment that characterized K-12 education, but it was a concern of assessors in the professions as well, who worried that by making their assessments authentic, they made them irreproducible and therefore unreliable. Forty plus years after the arrival of authenticity on the professional assessment scene, the discussion has changed. Rigorous investigation into assessment techniques in medical education, in particular, has demonstrated that the authenticity tradeoff as it was originally argued is a fallacious one. Medical educators have discovered a variety of ways to imbue authentic assessments with reliability, and vice versa. This chapter discusses the historical discussion around authenticity, and looks closely at three signatory assessments in medical education to glean lessons for assessors in other professions in bridging this supposed divide.

Takeaways

- Authenticity, the degree to which an assessment mirrors the ways in which tested knowledge and skills will be used in the real world, is a critical characteristic of assessment in the professions.
- The historical dialog around authentic assessment presented a supposed tradeoff between the validity imbued by authenticity and the reliability lost when assessments were made more complex and idiosyncratic to each learner and test experience.
- Considerable work on assessment in medical education has demonstrated numerous ways to maintain reliability while increasing the authenticity of assessments. Specific examples are discussed herein.

C. O'Neal (✉)
Faculty of Education, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
e-mail: coneal@mednet.ucla.edu

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_4

At a school in the American Midwest, a group of law students participated in a mock trial. The group has been separated into prosecution and defense teams while faculty member played the role of judge. Actors played the role of witnesses, the defendant, and the jury. Their professor made careful note of their arguments, and read through the extensive briefs they had prepared. Their grade for the assessment did not depend on who wins or loses the case, but on how sound their legal argumentation was, and how well presented.

Across the globe, a trained observer watches a second year medical student perform in an Objective Structured Clinical Examination (OSCE), noting how well the student interacts with an actor role-playing a case of angina, and giving marks to the student for each point she hits in a checklist of dealing with such patients.

And in still another part of the world, a team of environmental engineering students take notes from a "client." The client is a local halfway house and the students have been "hired" to plan and execute a pond and greenspace for the occupants of the house. The students' instructor will base their grade on feedback from the client, the soundness of their plan, and the quality of their finished product.

4.1 Introduction

Each of the above cases presents an example of learner assessment in professional education. Unlike more traditional forms of assessment, each of these "tests" has been crafted with an eye toward the *authenticity* of the experience, or the degree to which the assessment mimics the real-world scenarios in which learners will be expected to eventually perform.

For most educators in the professions, authenticity in education (and even assessment) is a given; indeed one could reasonably propose that authenticity is the *sine qua non* of assessment in professional education, and schools that ignore this imperative lose much credibility in their certification of future professionals. As a term, authenticity has become such a standard part of our assessment of professional learners that it has all but disappeared from the literature, and more recent searches on the term "authentic assessment" turn up far fewer returns than they did ten years ago. Authenticity has been subsumed into the more generic psychometric term of validity and is considered an indispensable characteristic when evaluating an assessment's value.

The term authenticity, as it is applied to assessment, has varied considerably in meaning since its inception in the early 1980s.[1] Wiggins (1998), in an attempt to

[1] As with many educational concepts, authentic assessment is poorly and inconsistently defined in the literature (Frey et al. 2012). It has been, and continues to be, conflated with "performance assessment" and "formative assessment" (Baker and O'Neil 1996). This is an understandable development of usage, as the three movements arose from similar motivations. But it is important to remember that performance can be independent of context, whereas authentic assessment is always crafted with an eye towards the real world context of implementation (Wiggins 1991) and formative assessment more accurately describes the intended use of the assessment rather than what is being assessed.

place some boundaries around a slippery subject, offered up a satisfying definition of authentic assessment containing six essential characteristics:

1. The assessment experience reflects the way content, skills, and behaviors are implemented in the real world.
2. The assessment requires the learner to make a series of informed choices in order to navigate a problem with many potential outcomes.
3. The assessment requires action on the part of the learner, and those actions would be recognizable to an expert as inherit to the field being tested.
4. The context of the assessment is as similar as possible to the context of the real-world equivalent.
5. The assessment requires the learner to employ a range of complementary skills in order to navigate the problem.
6. The assessment includes feedback on performance, and the opportunity to be reassessed after having incorporated that feedback. Because of this, authentic assessment sometimes overlaps with "performance assessment."

Wiggins' definition of authentic assessment can be seen as too constraining in its scope, but it is the most commonly cited definition and will serve us well for the purposes of this chapter.

Authenticity has not always been a given in professional education assessment. During much of the 1980s and 1990s, a fierce argument raged over the relative value of authenticity in assessment. Figure 4.1 summarizes the competing concerns regarding authenticity that predominated at the end of the past century. In this simplified model,[2] more inauthentic assessments were seen as more likely to suffer from "validity risk," where performance on the assessment would not accurately describe a learner's readiness to utilize those skills and knowledge in the workforce. On the other end of the continuum, truly authentic assessments were expected to deliver a highly valid measure of a learner's ability to perform, but were likely to suffer considerably from issues of "reliability risk." While this relationship is a truly oversimplified one, it does capture the argument of the time in the broadest of brush strokes (Gipps 1995).

In this chapter, we will revisit this classic argument. With 20/20 hindsight we will look at the perceived tradeoffs that have framed modern assessment in the professions and ask what we value in assessment. How should we assess our own assessments? Using cases from medical education, I hope to show that this historical argument was a valid one, but one misguided by the inappropriate

[2] It is important to note that Fig. 4.1 describes a simplified state of relative validity and reliability risk. As we will see later in this chapter, there is no reason that an inauthentic assessment could not be made highly valid, nor is there any reason that a truly authentic assessment could not be made highly reliable. But when comparing two assessments at either end of the continuum the difference in relative risks of invalidity and unreliability are worth addressing. Additionally, it is important to note that this model lumps various types of validity together, but is probably most descriptive of content and construct validity over other descriptors of validity.

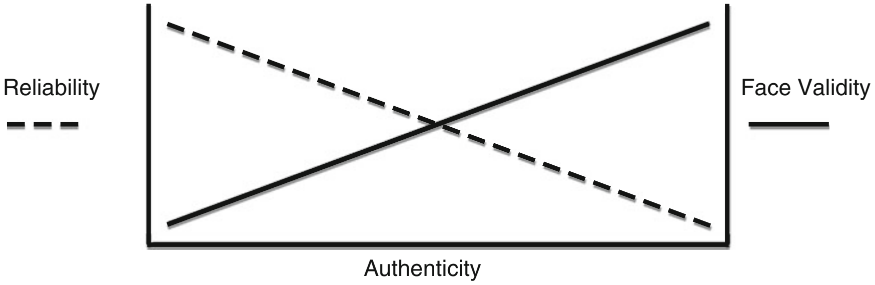


Fig. 4.1 Relative risks of an assessment being invalid or unreliable in relation to its authenticity

application of constraints. By looking at how contemporary medical educators have created assessments that maximize both reliability and validity, I hope to offer guidance to other disciplines still perfecting their application of assessment.

4.2 The Arguments for and Against Authentic Assessment in Professional Education—A Historical Perspective

Professional education reformers began to cry foul on traditional assessments as early as the 1980s and 1990s in both the U.S. and Europe[3] (Wiggins 1991; Archbald and Newman 1988). As concerns rose about the "product" that professional schools were delivering, the standard assessment toolbox of tests and quizzes quickly became a target of reformers looking to explain the disconnect between the hundreds of hours that went into a western professional education and the poor preparation of those students to succeed in a modern workforce. Traditional modes of assessment appeared to have low predictive success for professionals upon graduation (Darling-Hammond and Snyder 2000), and did not seem to reinforce the types of knowledge needed to succeed in the professions (Gibbons et al. 1994). Authentic assessment, proponents argued, measured the actual criterion that we were aiming for in performance, while less authentic assessments targeted imperfect indicators of that performance; it was, ultimately, an issue of validity.

The validity of an assessment describes how accurate inferences about future competence are that are based on performance on that assessment (Messick 1989), whereas reliability or generalizability describes the variability in assessment scores

[3] Indeed, the roots of this movement run as far back as the 1950s, with Lindquist (1951; p. 152) arguing that "it should always be the fundamental goal of the achievement test constructor to make the elements of his test series as nearly equivalent to, or as much like, the elements of the criterion series as consequences of efficiency, comparability, economy, and expediency will permit." (quote found by this author in Linn et al. 1991).

on that same assessment due to error (Norcini and McKinley 2007).[4] For most of the history of educational assessment thinking, a premium had been placed on the reliability of assessments, even to the disadvantage of validity (Broadfoot 1996; Resnick and Resick 1992; Linn et al. 1991). This lean toward reliability was motivated in no small part by the desire for comparability between learners. Employers want to be able to reliably distinguish between the candidates educators are producing. Assessments to this day, driven by this focus on reliability, have tended to assess what was most easily measured, whether it reflected the potential for application to its intended career or not.

But as early as the 1970s and 1980s, the dialogue around assessment began to describe conventional tests as examples of "Sign Assessment" (the measurement of items *related* to a behavior or construct), and to point out their obvious lack of validity when compared to "Sample Assessment," (the measurement of the behavior or construct itself; Wolf 1995). The attack on *reliability myopia* that began at this time may be considered the earliest rumblings of the authenticity movement. Authenticity proponents questioned the value of reliable tests that told assessors little about a learner's potential to succeed in the workplace. Content validity and face validity, among other measures of authenticity, began to gain traction as desirable, characteristics of professional education assessment. In one field, medical education, this period saw the rise of many of the more authentic assessments in use today: standardized patient assessments (Vu et al. 1987), Simulations (Norman et al. 1985), and Objective Structured Clinical Exams (Harden 1988) all gained early tracking during this renaissance of authenticity. Other professional disciplines soon followed suit (Darling-Hammond et al. 1995).

Proponents of authentic assessment did not see this argument as simply a psychometric one. If assessment really does drive the curriculum, then inauthentic assessments, they argued, had the potential to drive our curriculum and pedagogy away from the skills that really mattered to focus on imperfect indicators of performance that may or not translate into actual performance (Linn et al. 1991). Authenticity in assessment, on the other hand, was upheld by many as the guiding light that would refocus the curriculum and educators toward producing workers and professionals who could thrive in their chosen vocation. And this claim appears to have been largely true, with some claiming that the rush to authentic assessment that began in the 1980s has been the overriding driver of curricular and assessment reform in professional schools ever since (Linn et al. 1991).

[4] Following Gipps (1995), I use reliability in "relation to consistency as a basis for comparability; issues of consistent and comparable administration, comparability of the task, and comparability of assessment of performance (among raters)… rather than technical test-retest or split-half measures of reliability." Likewise, rather than parse validity into differing measures of construct, content, and criterion-related validity, I will instead use validity in its most general application of how well the test or measure in question is subjectively viewed to cover the concept it is claiming to measure, so called face validity. For an exceptional overview of the technical aspects of validity as they relate to authentic/performance assessment, I turn the reader to Moss (1992); additionally, Linn et al. (1991) broaden the consideration of assessment beyond reliability and validity in ways that are illuminating but beyond the aims of this chapter.

Other arguments in favor of authenticity were held up to convince educators and students of their value. For instructors in professional skills, what could be more satisfying than a test that accurately measures your students' preparation to put into action the knowledge and skills you are teaching them (Baron and Boschee 1995)? And from an educational theory viewpoint, the contextualized nature of authentic assessment jibes very nicely with social constructivist theories of how students learn; deeply understanding the material, representing it in a contextualized, nuanced way, and employing knowledge to affect meaningful action are all necessities for success on authentic professional assessments, and all are necessities of true learning according to constructivists (Maclellan 2004). Likewise, the holistic approach to assessment that characterizes most authentic assessment is seen as critical given the complex way that skills and knowledge are employed in real life (Wolf 1995).

Still other educational theorists argued that more authentic assessments would be, by necessity, more content specific. They argued that more content specific instruction and assessment was necessary given the poor evidence for the transferability of knowledge and skills from one domain to another. In the classic study of this phenomenon in medicine, Elstein et al. (1978) found that novice and expert clinicians differed little in the process they used to solve medical problems. Rather, the superior ability of medical experts to solve problems came from the broader, richer knowledge base that they could bring to bear on problems.[5] A more authentic assessment would therefore do a better job of distinguishing novices from experts, educational theorists argued, because it would rely as much on the context nuances of the real-life problem being presented as it would on logic and reasoning ability.

For learners in the professions, authentic assessments communicate the value of what they are learning in a way that builds intrinsic motivation and encourages them to engage more deeply with the material of the course (Maclellan 2004; Gibbs 1999; Newmann and Archbald 1992); to put it simply, they help make learning relevant and fun. Additionally, more authentic assessments were presupposed to do a better job of motivating students excited about finally getting to implement the skills they have been training to use (Baron and Boschee 1995; Svinicki 2004). They were also seen to improve learning transfer, implying that students are more likely to retain skills and content that they have already seen a compelling need for during training. Likewise, authentic assessment was thought to provide students with more credible feedback about their progress and how well they were being trained for the real world.

Despite authenticity's many admirable qualities, we must note that the call for more valid and authentic assessments was not without its critics (Baron and Boschee 1995). The primary concern, as might be expected, was about the

[5] Note that more recent analyses in the field of medicine, such as those done by Wimmers et al. (2007) suggest that content specificity alone does not completely explain differences in performance in the clinic. There is some X-factor that is independent to each learner that we must consider as well and that X-factor is likely to be some generalizable skill that each learner possesses to a greater or lesser degree.

reliability and generalizability of more authentic assessments (Hodkinson 1991; Prais 1991; Baron and Boschee 1995). Yes, authentic assessments might be more valid than traditional multiple choice tests, but if they were not also reliable, than that validity was for naught. And there is no doubt that examples of unreliable authentic assessments abounded at the time (e.g., the ubiquitous oral exam). In one classic example, Wolf and Silver (1986) were able to demonstrate considerable inter-rater variability for proctors evaluating even simple authentic tasks (in one case, evaluating the correct use of a micrometer; in another, evaluating the filling out of invoices). Even when time and energy were put into training evaluators to be more reliable, educational researchers were able to fairly easily demonstrate the persistence of strong inter-rater unreliability (Gipps 1995; Clark and Wolf 1991; Black et al. 1989).

Still others have argued that defining authentic assessment by its "real world" focus unfairly demeans more "traditional" pen and paper assessments and implies that they are somehow "inauthentic" (Terwilliger 1998). This is not necessarily the case as, particularly in the area of pure knowledge needed to succeed in a discipline, traditional testing might actually be a fairly "authentic" way to test. For example, taxi drivers in London are expected to pass a grueling written and oral exam on navigating London's streets. The test, with its emphasis on recall of routes and landmarks, can be seen as a fairly authentic test of the way the knowledge would be used on the job. These same critics also pointed out that validity is not necessarily a characteristic of an individual assessment, but is instead a descriptor of how that assessment is interpreted and used (Frey et al. 2012). This critique particularly makes sense given the extreme variability in how assessments are implemented from one school to another, or even one class to another.

In response to these attacks, defenders of authentic assessment leapt to their standard's defense, claiming that, yes, reliability is critical in norm-referenced grading, but if authentic assessments are criterion-referenced, and established according to an external reference point, then reliability becomes moot as we are not comparing learner to learner, but rather, learner to standard (Burke and Jessup 1990). For Wiggins—arguably the godfather of authentic assessment—the challenges of unreliability inherent in authentic assessment are acknowledged, but only insofar as they are seen as hurdle that can and must be leapt (Wiggins 1993, 1998). One cannot help but enjoy the fervor of Wiggins' argument as he upholds the primacy of authenticity as the driver of assessment reform efforts, and the employment of easier and more reliable assessments as a betrayal most foul of that reform.

With this argument in full sway, critics of authentic assessments were keen to highlight other ways in which these newer assessments fell short. Even when they acknowledged the increased validity of these assessments, they were quick to point out that they were more highly specific than originally envisaged, with performance on one set of tasks not very predictive of performance on even closely related but distinct tasks (Greeno 1989). Just because a medical student performs well diagnosing a simulated patient with chronic obstructive pulmonary disease, we cannot assume they will perform equally well when asked to diagnose a related but distinct entity like emphysema. This was particularly damning from the point of view of the critics of authentic assessment, who were already concerned at the considerable resources going into a time-intensive activity that appeared to be highly context-dependent with poor generalizability.

Beyond these psychometric concerns, authentic assessment is not without its additional challenges. Svinicki (2004) emphasizes the unavoidably time intensive nature of preparing and participating in authentic assessments, both for the student and instructor. A single Objective Structured Clinical Exam (like the OSCE described in the example that started this chapter) can take thousands of dollars and hundreds of people hours to design and implement, an investment that demands real value in the data it returns. Even when the time and money is invested, those expenditures mean less overall time for more formative assessments from which learners might benefit, and less time for assessors to network between themselves (Wolf 1995).

Svinicki also points out that if students are to get the most formative returns out of the assessment they must actually perceive it as accurately authentic. Gulikers et al. (2008) found that students perceived far less utility in an assessment they deemed inauthentic and considered it as less valuable than other assessments. One could surmise that those students were therefore less likely to take any lessons learned from the less authentic assessment to heart. Indeed, novice learners are not always the best qualified to judge an assessment's authenticity and their refusal to "play along" with an assessment that might actually hold value can sabotage the whole experience.

There are other less obvious concerns with implementing authentic assessment (E.g., safety and supervision issues arise when authentic assessments blur the line between student and worker, as in the engineering design team example at the beginning of this chapter), but most of the literature around authentic assessment has eschewed these more esoteric concerns for the meat and potatoes issues of authentic assessment's validity and reliability, and it is on these thorny subjects that the remainder of this chapter will focus. I have also provided a summary of the traditionally perceived strengths and weaknesses of authentic assessments in Table 4.1.

Table 4.1 A summary of the historical argument around authentic assessment

Strengths of authentic assessment	Weaknesses of authentic assessment
More accurately reflects how the learner is likely to perform in the "real world"	More difficult to compare a learner's performance to others' performance
Focuses instruction/curricula on the "things that matter"	Expensive to create and administer, both in terms of money and time
Better reflects the social constructivist view of how students learn	Different raters are more likely to assess the same learner differently on more complex, authentic assessments
Builds motivation in learners who see how what they are learning will eventually be used in the workforce	Highly specific tasks on assessments may not be very predictive of a learner's performance on a related but different assessment

It is very important to point out that authenticity is not a binary state wherein we can divide traditional exams as "inauthentic" and real-world-focused assessments as "authentic." Instead, all assessments exist on a continuum of authenticity (Wiggins 1998), such as that represented in Fig. 4.1. Neither can we use authenticity as the only measure of an assessment's worth. We might place the three cases that opened this chapter on that continuum ranging from least authentic to most authentic, but that ranking does not give us the full picture of the relative utility or appropriateness of each of these assessments.

Much of the argument around authentic education occurred in the context of large-scale assessment in the K-12 school systems of the U.S., Canada, and Britain. In an era when calls for authenticity were coinciding with reductions in resources, the argument over how best to spend precious assessment dollars became a passionate one. But the same argument was happening for different reasons in professional schools on both sides of the Atlantic, and in medical education in the U.S. particularly. The 1990s saw increased scrutiny of medical education in the U.S. as the public consciousness around poor health care outcomes, increased healthcare spending, and medical malfeasance became galvanized. Contemporaneously, psychologist Miller (1990) captured the medical education community's concerns about inauthentic assessment in his framework for assessing clinical competence. The (now well-known) pyramid placed *Knowing* at the bottom of the assessment framework, rightfully implying that the bulk of our assessment of medical learners occurred at this level. *Knowing how* and *Showing how* were each steps of competence and performance, finally reaching *Doing* at the top of the pyramid, again implying that we were doing a poor job of assessing competence at the level of actual performance.

It is likely not accidental that Miller's framework is essentially an argument for authentic assessment in medicine, and medical schools across the country responded to his critique by redoubling their efforts to improve assessment of their learners at all levels. This effort, and the efforts in quality improvement that preceded it make medical education a natural experiment of sorts to evaluate the argument around authenticity. Of all the professional disciplines in the U.S., medical education has faced the most pressure for reform, largely motivated by the issues described in the paragraphs above. It is my hope that by using the state-of-the-art in assessment to reflect on the historic argument for and against authenticity, we will generate some lessons for the other professional fields whose assessment philosophies continue to evolve.

4.3 Authentic Assessment in Medicine

Medical education is a highly connected field, with innovations that work spreading rapidly from school to school (and innovations that do not work usually spreading just as quickly). It is also a highly regulated discipline, with the National Board of Medical Examiners overseeing medical student certification, and the Accreditation

Council for Graduate Medical Education monitoring post-graduate medical education and the certification of residents. Both of these factors allow us to speak about the dominant assessment strategies in the field with considerable confidence that we are describing a sizable percentge of medical institutions. Local innovation does happen, frequently, but most schools use some variant of the assessment strategies described below.

For the purposes of breadth, we will consider three assessment methods that lie on different points of the authenticity spectrum; ranging from least authentic to most: *NBME-style multiple choice questions* for assessing didactically-delivered knowledge; *Objective Structured Clinical Examinations* for assessing junior medical trainees mastery of simple clinical skills; and the *Clinical Evaluation Exercise* for assessing performance at work in the clinic.

With each assessment, we'll explore their documented reliability and face validity, and assess whether the simple model in Fig. 4.1 holds up. We'll also ask what lessons have been learned in medicine that may be of use to other professional disciplines.

National Board of Medical Examiners (NBME)-style Multiple Choice Questions. Assessment of medical students' pre-clinical content knowledge remains firmly rooted in the use of multiple choice questions (MCQs) at most medical schools. This usage persists despite long established recognition that MCQs, as traditionally written, lack validity when assessing higher-order skills (Levine et al. 1970). MCQs in their traditional format represent the classic case of an assessment sitting at the very left of the authenticity spectrum: supposedly highly reliable, but with low face validity; in other words, inauthentic.

However, rather than eject MCQs from their repertoire based on this implied inauthenticity, medical assessors have sought methods for maintaining the reliability of these tests while, at the same time, pushing them up the validity curve to make them more authentic. This effort has been largely manifested in the use of so called "NBME-style" questions. The National Board of Medical Examiners (NBME) has issued considerable guidance on item writing for multiple choice testing in medical school. NBME-style questions are characterized by:

- One best answer formats (as opposed to true-false formats)
- An item stem that focuses on the application of a distinct clinical or biological concept
- Homogenous distractors (incorrect responses), and
- The absence of technical item flaws that give away correct or incorrect answers

While most of these characteristics serve to maintain the reliability of the test items, NBME item stems also strive for at least some authenticity by focusing on the clinical application of concepts learned in the pre-clinical years. The following is an example:

> A 34-year-old woman has had severe watery diarrhea for the past four days. Two months earlier she had infectious mononucleosis. She abuses drugs intravenously and has antibodies to HIV in her blood. Physical examination shows dehydration and marked muscle weakness.[6]
>
> Laboratory studies are most likely to show…
>
> A. decreased serum K^+ concentration
> B. decreased serum Ca^{2+} concentration
> C. increased serum HCO_{3-} concentrations
> D. increased serum Na^+ concentration
> E. increased serum pH

While the item tests students' understanding of homeostasis, it strives for some measure of authenticity by situating that knowledge in a commonly-encountered clinical scenario. Indeed, the scenario is a common enough one in hospital wards that the item stem could have been pulled off a real patient's chart.

The Reliability and Validity of NBME-style Multiple Choice Questions. MCQs have long been the mainstay of standardized assessments intended to measure and rank performance with a high degree of reliability. There is a reason that qualifying tests the world over predominantly rely on MCQs; they allow assessors to reliably measure performance across a variety of settings and learners. However, as we discussed earlier in this chapter, the validity of MCQs has long been suspect, and from the early days of the authentic assessment movement, they were held up as the inauthentic bogeyman of assessment.

NBME-style MCQs represent one of the best attempts to infuse MCQs with more validity. By situating the questions within a clinically-relevant context, the thinking is that the question will better assess how the learner might apply that knowledge in a real-world situation. There are two questions that can be asked about any test that supply insight into its validity: (1) does learner progress on the test increase over time? In medical education, we can answer this question by comparing the performance of novice students, advanced students, and residents on the same test, and (2) is performance on the test predictive of future performance in the workforce?

It turns out that, in general, performance on NBME-style MCQ tests does improve over time. In one study, medical student performance on the Comprehensive Basic Science Examination (a practice test for the USML Step 1 exam) improved linearly over time (Kibble et al. 2014). This correlation of experience to performance on such tests is seen in other studies as well and is heartening (Glew et al. 1997). It appears that, in terms of construct validity, at least, NBME-style questions can be seen as valid. Norman et al. (1987) also showed that there is a strong relationship between the content assessment of MCQ style tests and more writing-intensive Modified Essay Question tests MEQs), suggesting that MCQs are at least as valid.

[6] Example downloaded from http://medicine.tufts.edu/~/media/TUSM/MD/PDFs/Education/OEA/Faculty%20Development/Evaluation_Writing%20Exam%20Questions%20for%20Basic%20Sciences.pdf on December 17, 2015.

The degree to which NBME-style MCQs are predictive of future clinical performance is more challenging to measure and more contentious. The actual workforce performance of medical professionals is itself difficult to assess as so many providers tend to provide care for a single individual, and there are so many idiosyncrasies unique to each clinical situation. Despite this, there are studies that have attempted to correlate performance on a NBME-style exam and some measure of clinical performance. One study, for example, found a modest but significant correlation between the Family Medicine Shelf exam (which uses NBME-style MCQs) and clinical clerkship evaluations of students (Myles and Galvez-Myles 2003). While findings are not uniform, and the relationship is often weak, even if significant, other studies have confirmed an association between NBME-style tests and measures of clinical performance like the Objective Structure Clinical Examination (Simon et al. 2002).

Lessons from NBME-style questions for the Non-medical Professions. Assessors in medical education have reliably demonstrated that MCQs, long paragons of reliability, can also be infused with greater face-validity despite their apparent lack of authenticity. Given this, there is much that can be assessed with well-constructed MCQ's without ever having to set foot outside the testing hall, and there seems little to gain by ignoring this. Success in medicine depends on a hard-to-define mix of knowledge, skill, and affect, and a good physician has admirable qualities in all these areas. We could conceivably craft every assessment to a degree of authenticity that captures all of these qualities at once, but there may also be value in assessing one area separately from another, and MCQs allow us to assess a medical learner's fund of knowledge very well.

The ubiquity of MCQ-style exams in medicine reminds us that a broader view of authenticity may be needed. A multiple choice exam, one could argue, is authentic if it helps prepare the learner for future assessments. Since medical learners will take one MCQ exam after another (Step 1 and 2 exams, boards, and then routine recertification for years to come), there is considerable value to the learner in gaining early proficiency in succeeding on such tests. In this case there is a certain realpolitik in play that medical educators ignore at their peril.

We should also note that more valid, high quality NBME-style MCQs do not write themselves. The writing of such questions is a skill that requires training, practice, and monitoring for quality. Indeed, faculty who do not get explicitly trained on how to write high quality questions produce questions of questionable value and validity (Jozefowicz et al. 2002). Educators in other professions attempting to infuse their written assessments with more validity would do well to examine the training in item writing that has become expected in medicine (Case and Swanson 1998).

4.4 Objective Structured Clinical Exams

The OSCE represents one of medical education's first attempts to infuse assessment methodologies with authenticity (Harden and Gleeson 1979). In the OSCE, medical students perform discrete clinical tasks, sometimes as defined as interpreting an X-ray, sometimes as complex and ill-defined as interviewing and examining actors trained to accurately and consistently portray a patient with a disease, so called "Standardized Patients" (Khan et al. 2013). Ideally, each student rotates through a panel of clinical tasks, demonstrating 20–40 skills in a few hours. While OSCEs can not assess performance in all situations (e.g. life threatening clinical situations are more challenging to simulate in this way), they are seen as particularly useful for assessing student performance in handling a variety of clinical conditions and tasks. At first glance, OSCEs would appear to fall somewhere in the middle of the authenticity spectrum, and therefore be susceptible to both validity risk and reliability risk.

The Reliability and Validity of Objective Structured Clinical Exams. Few assessment tools in medicine (or any discipline for that matter) have been studied as much as OSCEs. A PubMed search for just the term, "OSCE," returns 843 articles, a sizeable chunk of which address the psychometric aspects of the assessment, at least in part.

OSCEs appear to be reasonably valid tool for assessing clinical competence, depending in some part on how validity is defined. Multiple authors have demonstrated the validity of OSCEs in different ways, some showing that more experienced practicioners perform better on the assessment than more junior ones (Cohen et al. 1990). Others, likewise, have found correlations between OSCEs and later tests of medical knowledge and skills, including USMLE Step 2 and 3 scores (Dong et al. 2014). Reported reliability of OSCEs has been more varied, with authors like Cohen et al. (1990) reporting impressive reliability coefficients in the 0.80–0.95 range (for OSCEs ranging from 15 to 50 stations). On a purely psychometric level, both of these measures are influenced by the length of the OSCE, the number of different skills being assessed, and the amount of time devoted to each station (Pell et al. 2010; Norcini and McKinley 2007). Most authors agree with Epstein (2007) that an OSCE of 14–18 stations and 5–10 min allowed for each station generates reasonable reliability measures.

Beyond the length and structure of the OSCE, three factors appear to offer assessors the most opportunity to positively influence the reliability of the assessment: the use of standardized scoring rubrics (Smee 2003), the use of trained examiners (Vleuten et al. 1989), and the use of highly-trained standardized patient performers (Smee 2003). When one of these factors is missing or poorly realized, the reliability of the assessment suffers, but likewise, when care is put into the development and training of performers and assessors, OSCEs can be made highly reliable (Norcini and McKinley 2007).

Part of the success of OSCEs, and their ubiquity in medical education, can be traced to the way they are assessed. When first conceived, and ever since, OSCEs have depended on a mix of checklist and rating scale. The use of both of these types of evaluation methods was seen as improving the objectivity, and therefore the reliability of the tool. This may not be true, with multiple authors having shown that ratings alone may be just as reliable as the theoretically more objective checklists (Cunnington et al. 1997).

Of the various ways to introduce greater reliability into an authentic assessment, few have received as much attention as rubrics. Rubrics "*consist of a fixed scale and a list of characteristics describing performance, the provide important information to teachers…and others interested in what students know and can do. Rubrics also promote learning by offering clear performance targets to students for agreed-upon results*" (from Marzano et al. 1993 quoted on page 53 of Baron and Boschee 1995).

Lessons from OSCEs for the Non-medical Professions. As a measure of performance in a simulated setting, OSCEs fall somewhere on the "more authentic" side of the authenticity scale. Historical critics of authentic assessments would likely worry whether the increased risk of unreliability in this assessment is outweighed by the benefits of the increased validity. In reality, OSCEs, perhaps more than any other assessment have shown us that you can have your cake and eat it too, for a price. The price with OSCEs is in training time for assessors and standardized patients who are critical for the performance of the exam. If considerable time and effort is put into the development and training of each, reliability measures for OSCEs can rival those of far less authentic assessment tools.

This cost in training and development is quite high, however, and more poorly resourced disciplines and institutions may find the burden of implementing an OSCE-like assessment with their students simply too expensive. Medical schools have offered a potential solution to this in the pooling and sharing of OSCE resources. In multiple states, medical schools share OSCE scripts and training tools with each other, meaning that no one school has to take the time and energy to independently develop those tools. Likewise, institutions will share the psychometric characteristics of the OSCEs they perform, letting the community as a whole promote sounder OSCEs and get rid of more problematic ones.

As with other assessment tools, with OSCEs we must be careful what we are testing. Just because we watch a student place their stethoscope in all the right spots during a physical exam, we can not then say that that student knows how to interpret the results of that exam. OSCEs are only useful in so far as we collect from students the data that informs us of their performance. OSCEs themselves should be seen as measuring performance in a simulated environment, the leap from that performance to competence is one the assessor must make very carefully (Khan and Ramachandran 2012).

4.5 Clinical Evaluation Exercises

Since the mid-1970s, the Clinical Evaluation Exercises (CEX) has been used to provide assessment of a medical learners in the actual clinical setting during a period of actual clinical practice. A common CEX scenario consists of an experienced preceptor observing a medical resident take a history and perform a physical on a new patient. Whereas the OSCE measures performance in a simulated setting, the CEX occurs during actual practice and may therefore be seen as deeply authentic.

Similarly to the OSCE, the CEX is onerous to perform. Each session may take up to two hours per learner. Because of this, an alternative mode called the mini-CEX has evolved that limits the observation and feedback period between resident and preceptor to 15–20 min. This more compact format allows for more sessions to occur for any given learner during the year, and for more varied clinical situations to be assessed during any given assessment period. In terms of authenticity, Norcini et al. (2003) argue that the mini-CEX is even more authentic than it's longer cousin, the CEX, as the mini-CEX mimics the time-crunched nature of most clinical encounters, whereas the traditional CEX, with its hours long process, is an artificial setting compared to most clinics.

The Reliability and Validity of Clinical Evaluation Exercises. Falling as it does on the more authentic side of the authenticity scale, we might predict that the CEX is highly valid, but problematically unreliable. And indeed, the CEX is considered a very valid measure of medical learner clinical competence, both because it is an assessment that happens in situ, and because performance on the CEX improves with greater experience and is predictive of later clinical performance (Al Ansari et al. 2013). For medical education, the CEX provides an unprecedented opportunity to see how learners are mastering the skills they will actually use in the work force.

When we look at the potential reliability of the CEX, we do indeed see that is fraught with threats to reliability. Since patients vary from clinic to clinic and even from day to day, it is virtually impossible to provide each learner with an identical assessment experience. Furthermore, because of the time-intensive nature of the CEX, assessment is typically performed by a single faculty member who's assessment of performance may or may not align with that of her peers'. This lack of generalizability fueled early worry about the CEX as an assessment tool (Kroboth et al. 1992; Norcini 2005; Norcini et al. 2003) and must be taken into consideration when implementing CEXs.

Compared to the CEX, the mini-CEX actually appears to return satisfactory reliability numbers when enough sessions occur. Just as with OSCEs, repetition can be used to drive up reliability numbers. In one study (Norcini et al. 1995), ten mini-CEX sessions were required to produce a confidence interval of ± 0.46 on a five point scale. Increasing the number of sessions beyond ten resulted in only marginal gains in confidence. Incidentally, this same study asked preceptors to rate their satisfaction with the mini-CEX as an assessment tool; the average satisfaction rating was 7.0 ± 1.3 out of 9 (with 9 being the most satisfied). This is a fuzzy measure of validity, but lends more credence to the validity of the tool.

Lessons from Clinical Evaluation Exercises for the Non-medical Professions. Imbuing truly authentic assessments like the CEX with high degrees of reliability relies on tightly linking assessors in via discussion and networking to achieve norming of evaluation and grading (Wolf 1995). Regular standardization of grading practices, communication between assessors at multiple sites, and continuous discussion of the quirks of any given assessment tool are all methods by which reliability may be driven up for authentic assessments.

As with the OSCE, the introduction of structured assessment forms can improve the reliability of the assessment. Winckel et al. (1994) were able to generate inter-rater reliabilities as high as 0.78 by introducing a structured technical skills assessment form into the clinical evaluation of surgery procedures.

Surprisingly enough, while scoring crutches like rubrics can help (Shavelson et al. 1992), the clarity of standards established to grade authentic assessments do not appear to be nearly as important as the discussion of those standards that happens when assessors interconnect (Wolf 1995). It seems that no standard can be well-written enough to prevent inter-rater differences of interpretation. Only by investing the time in discussion can those standards be truly understood by the community of assessors (Gipps et al. 1991). Various modes have been utilized to generate this discussion, including face-to-face evaluation meetings or online discussion. All are time consuming, but it is important to remember that there are likely to be multiple benefits from time invested in such discussions. Assessments will be done more reliably, assessors will feel more unified with others on their team, and issues of instruction are likely to be discussed, in addition to other benefits.

Finally, as with OSCEs, the importance of "multiple validations" can not be overstated as key to improving the reliability of highly authentic assessments. While one clinical evaluation is likely to describe little more than the learner's performance in that discrete task, over time the accumulation of evaluations begins to tell a complete picture of that learner's overall performance and potential (Baron and Boschee 1995).

4.6 Conclusion

What should we, as educators in the professions, value in the assessments we create and use? We began this chapter by revisiting the argument around this question that divided educational assessors during the latter two decades of the past century. For that group, the supposed tradeoff between validity and reliability in authentic assessment seemed a reasonable one to discuss and parse, but as the science of assessment in the professions has evolved, it has become clear that we can neglect neither. The stakes in professional education are too high, the consequences of poor assessment too dire. The discussion must not be about favoring reliability over validity; it must be about creating the most authentic assessments we can that are also as reliable as they can be.

Assessors in medical education have shown us that, with some effort, any assessment may be made reliable, no matter how authentic it may be. Likewise, seemingly inauthentic assessment modes may have increased validity imparted on them by focusing on the real-life applications of the basic knowledge being assessed. Disciplines that wish to follow in the footsteps of medicine have clear examples to follow. Could the increased authenticity that has been added to multiple choice questions in medical board exams be added to MCQs in engineering? Could the setting and style of the objective structured clinical exam be adapted to law? Might the principals of truly authentic assessment that embody the Clinical Evaluation Exercise be adapted to "on the job" evaluations in pharmacy? All of these techniques have obvious analogs in the other professional disciplines, and all represent realistic examples of maximized validity and reliability.

Perhaps more important than the individual value of any one assessment is the combined value of the assessment suite that an institution uses. Lee and Wimmers (2011) highlighted this finding in their comparison of three commonly used clinical clerkships assessments: preceptor's evaluations, OSCEs (like those described above), and board exams (exemplary of the NBME-style MCQ tests described above). They found that each of the three tools contributed a different part of the story of any given student's performance. Taken together, they clearly were commonly indicative of some factor of clinical competence. Each assessment alone was useful, yes, but together, they gave a richer, more complete picture of each student's abilities.

These are good lessons for professional educators to learn, but ultimately, the argument over validity and reliability of authentic assessment turns out to be the wrong one for professional education. That argument, framed as it was in the context of resource-poor K-12 education, was understandably stilted toward concerns over just how costly it would be to introduce reliable authentic assessment into high-throughput educational systems where frequency of assessment was high and available of assessors was limited. Professional schools do not have the luxury of such an excuse.

Van der Vleuten and Schuwirth (2005) proposed an expansion of the characteristics by which we might choose an assessment. In addition to the traditional descriptors of reliability and validity, they have added the concepts of educational effect, feasibility, and acceptability. *Educational effect* describes the catalyzing effect that an assessment might have in improving or even motivating students' learning of a skill or topic. *Feasibility* describes how affordable and implementable a particular assessment might be. This is closely related to, but distinct from *acceptability*, which describes how likely instructors and students are to endorse the particular assessment tool or method.

Both validity and reliability have become seen as self-evident necessities of assessment in medicine. The dialogue around design and choice of assessment has moved on from the historical argument and turned instead to these three variables. If any of these three is completely lacking, then the reliability and validity of an assessment become moot (Norcini and McKinley 2007). The three medical learner assessment methods described in this chapter have all been weighed on their educational effect, feasibility, and acceptability, and while issues exist with each, each has found support for its use.

Finally, Norcini et al. (2003), reminds us that the most authentic assessments occur not in school or training programs, but in the workplace. Because of the considerable overlap between the classroom and actual practice in the professions, professional educators would be wise to broaden their view of assessment beyond that of a practice that stops upon graduation. He suggests three tools for assessing performance in the medical workplace: outcomes measures (e.g., patient morbidity and mortality), large-scale data collection and processing (e.g., chart review), and portfolios. There is no reason that these three tools, and others, could not be implemented earlier in a professional's education. If they were, it might be the first step toward creating a truly authentic assessment system that began in the first year of professional school, but continued throughout a professional's career.

Issues/Questions for Reflection

- Despite the potential to construct assessments that are both highly authentic and highly reliable, what are the indispensible characteristics of assessment in the professions? Are there qualities of professional assessment that deliver the highest return on investment?
- Could K-12, higher education, and professional postsecondary education benefit from greater cross-level dialogue? For example, does the emphasis on authenticity at the professional level hold promise for better preparing K-12 students for what is to come?

References

Al Ansari, A., Ali, S. K., & Donnon, T. (2013). The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Academic Medicine, 88*(3), 468–474.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Washington DC: Office of Educational Research and Improvement.

Baker, E. L., & O'Neil Jr, H. F. (1996). Performance assessment and equity. *Implementing performance assessment: Promises, problems, and challenges*, 183–199.

Baron, M. A., & Boschee, F. (1995). *Authentic assessment: The key to unlocking student success*. Lancaster, PA: Order Department, Technomic Publishing Company, Inc.

Black, H., Hale, J., Martin, S., & Yates, J. (1989). *The quality of assessment*. Edinburgh: Scottish Council for Research in Education.

Broadfoot, P. (1996). *Education, assessment and society: A sociological analysis*. Open University Press.

Burke, J., & Jessup, G. (1990). Assessment in NVQs: Disentangling validity from reliability. *Assessment Debates*, 188–196.

Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences* (2nd ed.). Philadelphia, PA: National Board of Medical Examiners.

Clarke, L., & wolf, A. (1991). Blue Badge Guides: Assessment of national knowledge requirements. Final Project Report to the Department of Employment (unpublished).

Cohen, R., Reznick, R. K., Taylor, B. R., Provan, J., & Rothman, A. (1990). Reliability and validity of the Objective Structured Clinical Examination in assessing surgical residents. *The American Journal of Surgery, 160*, 302–305.

Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1997). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education, 1*, 227–233.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. Teachers College Press.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and teacher education, 16*(5), 523–545.

Dong, T., Swygert, K. A., Durning, S. J., Saguil, A., Gilliland, W. R., Cruess, D., et al. (2014). Validity evidence for medical school OSCEs: Associations with USMLE® step assessments. *Teaching and Learning in Medicine, 26*(4), 379–386.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press.

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356* (4), 387–396.

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation, 17*(2), 2.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage.

Gipps, C. (1995). Reliability, validity, and manageability in large-scale performance assessment. *Evaluating authentic assessment*, 105–123.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. *Assessment Matters in Higher Education*, 41–53.

Gipps, C., McCallum, B., McAlister, S., & Brown, M. (1991). National assessment at seven: some emerging themes. In C. Gipps (Ed.), *British Educational Research Association Annual Conference*.

Glew, R. H., Ripkey, D. R., & Swanson, D. B. (1997). Relationship between students' performances on the NBME Comprehensive Basic Science Examination and the USMLE Step 1: A longitudinal investigation at one school. *Academic Medicine, 72*(12), 1097–1102.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44*(2), 134.

Gulikers, J. T., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008). Authenticity is in the eye of the beholder: Student and teacher perceptions of assessment authenticity. *Journal of Vocational Education and Training, 60*(4), 401–412.

Harden, R. M. (1988). What is an OSCE? *Medical Teacher, 10*(1), 19–22.

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 12*, 41–54.

Hodkinson, P. (1991). NCVQ and the 16-19 curriculum. *British Journal of Education and Work, 4* (3), 25–38.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine, 77*(2), 156–161.

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher, 35*(9), e1447–e1463.

Khan, K., & Ramachandran, S. (2012). Conceptual framework for performance assessment: competency, competence and performance in the context of assessments in healthcare–deciphering the terminology. *Medical teacher*, *34*(11), 920–928.

Kibble, J. D., Johnson, T. R., Khalil, M. K., Peppler, R. D., & Davey, D. D. (2014). Use of the NBME Comprehensive Basic Science Exam as a progress test in the preclerkship curriculum of a new medical school. *Advances in Physiology Education, 38*, 315–320.

Kroboth, F. J., Hanusa, B. H., Parker, S., Coulehan, J. L., Kapoor, W. N., Brown, F. H., et al. (1992). The inter-rater reliability and internal consistency of a clinical evaluation exercise. *Journal of General Internal Medicine, 7*(2), 174–179.

Lee, M., & Wimmers, P. F. (2011). Clinical competence understood through the construct validity of three clerkship assessments. *Medical Education, 45*(8), 849–857.

Levine, H. G., McGuire, C. H., & Nattress Jr, L. W. (1970). The validity of multiple choice achievement tests as measures of competence in medicine. *American Educational Research Journal*, 69–82.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. *Educational Measurement*, 119–158.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Maclellan, E. (2004). Authenticity in assessment tasks: A heuristic exploration of academics' perceptions. *Higher Education Research & Development, 23*(1), 19–33.

Marzano, R. J., Pickering, D. J., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. Aurora, CO: McREL Institute.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington DC: Oryx Press.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229–258.

Myles, T., & Galvez-Myles, R. (2003). USMLE Step 1 and 2 scores correlate with family medicine clinical and examination scores. *Family Medicine-Kansas City-, 35*(7), 510–513.

Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. *Toward a new science of educational testing and assessment*, 71–83.

Norman, G. R., Smith, E. K. M., Powles, A. C. P., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21* (4), 297–304.

Norman, G. R., Muzzin, L. J., Williams, R. G., & Swanson, D. B. (1985). Simulation in health sciences education. *Journal of Instructional Development, 8*(1), 11–17.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476–481.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*(10), 795–799.

Norcini, J. J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical Education, 39*(9), 880–889.

Norcini, J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching and teacher education, 23*(3), 239–250.

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics-AMEE guide no. 49. *Medical Teacher, 32*(10), 802–811.

Prais, S. J. (1991). Vocational qualifications in Britain and Europe: theory and practice. *National Institute Economic Review, 136*(1), 86–92.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In *Changing assessments* (pp. 37–75). Netherlands: Springer.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, 22–27.

Simon, S. R., Volkan, K., Hamann, C., Duffey, C., & Fletcher, S. W. (2002). The relationship between second-year medical students' OSCE scores and USMLE Step 1 scores. *Medical Teacher, 24*(5), 535–539.

Smee, S. (2003). ABC of learning and teaching in medicine: skill based assessment. *BMJ: British Medical Journal, 326*(7391), 703.

Svinicki, M. D. (2004). Authentic assessment: Testing in reality. *New Directions for Teaching and Learning, 2004*(100), 23–29.

Terwilliger, J. S. (1998). Rejoinder: response to Wiggins and Newmann. *Educational Researcher, 27*(6), 22–23.

Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*(3), 309–317.

Vleuten, C. V. D., Luyk, S. V., Ballegooijen, A. V., & Swanson, D. B. (1989). Training and experience of examiners. *Medical Education, 23*(3), 290–296.

Vu, N. V., Steward, D. E., & Marcy, M. (1987). An assessment of the consistency and accuracy of standardized patients' simulations. *Academic Medicine, 62*(12), 1000–1002.

Wiggins, G. (1991). Teaching to the (authentic) test. *Educational Leadership, 46*, 41–47.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(3), 200–208.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance*. San Francisco, CA: Jossey-Bass Publishers. 94104.

Winckel, C. P., Reznick, R. K., Cohen, R., & Taylor, B. (1994). Reliability and construct validity of a structured technical skills assessment form. *The American Journal of Surgery, 167*(4), 423–427.

Wimmers, P. F., Splinter, T. A., Hancock, G. R., & Schmidt, H. G. (2007). Clinical competence: General ability or case-specific? *Advances in Health Sciences Education, 12*(3), 299–314.

Wolf, A. (1995). Authentic assessments in a competitive sector: Institutional prerequisites and cautionary tales. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment*. Open University (Cited).

Wolf, A., & Silver, R. (1986). Work based learning: Trainee assessment by supervisors.

Chapter 5
Assessing Performance in Engineering Education: Examples Across Fifty Years of Practice

John Heywood

Abstract This chapter outlines the challenges associated with defining assessment in the context of engineering education. A case study undertaken in the late 1950s brought into question the validity of written examinations. One response to this was broadening the concept of coursework to include projects and the application of criterion referenced measures to assessment. It is against this backdrop that discussion takes place about recent developments in outcomes (competency)-based assessment required by regulation agencies. Studies of engineers at work show the complexity of engineering and the significance of the affective domain and tacit knowledge in acquiring competence. The idea that an engineer should be able to take on a professional role immediately after leaving college without some prior guided experience of working in industry is refuted. This chapter supports the view that engineering education may better prepare students for work in the industry if it is structured on a cooperative basis. But the experience of industry must be carefully designed. The acquisition of competence occurs through a developmental process and is subject to the experience of the conditions imposed by the organization. This has implications for the way in which formative assessment is conducted.

Takeaways

- Studies of engineers at work show the complexity of the activity of engineering and the significance of the affective domain and tacit knowledge in acquiring competence.
- The idea that an engineer should be able to take on a professional role immediately on leaving college without some prior guided experience of working in industry is shown to be nonsense.

J. Heywood (✉)
Professorial Fellow Emeritus of Trinity College Dublin, University of Dublin, Dublin, Ireland
e-mail: heywoodj@eircom.net

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_5

- It seems clear that the acquisition of a competence is through a developmental process and subject to the experience of the conditions imposed on it by the organisation.

5.1 Introduction

'Assessment' seems to have become a term that means what you want it to mean. Sadler (2007) observed "that many of the terms we use in discussion on assessment and grading are used loosely. By this I mean we do not always clarify the several meanings a given term may take even in a given context, neither do we necessarily distinguish various terms from one another when they occur in different contexts. For example, the terms 'criteria' and 'standard' are often used interchangeably". Yokomoto and Bostwick (1999) chose 'criteria' and 'outcome' to illustrate problems posed by the statements in ABET, EC 2000 (ABET, 1997). In this text there is room for plenty of confusion between such terms as 'ability', 'appraisal', 'assessment', 'capability', 'competency', 'competences', 'competencies', 'criteria', 'criterion-referenced', 'evaluation', 'objectives', 'outcomes' and 'objectives', all of which will be used in the text that follows. Semantic confusion abounds and there is disagreement about plurals as for example 'competences' or 'competencies'.

McGuire (1993), a distinguished medical educator, argues that the use of the term performance is a naive disregard of plain English. We are not concerned with performance per se, "rather we are concerned about conclusions we can draw and the predictions we can make on the basis of that performance. And this is a high inference activity".

Assessment is now commonly used to describe any method of testing or examining that leads to credentialing, and as such embraces 'formative' assessment which while not contributing directly to credentials helps students improve their 'learning', and 'self-assessment' that may or may not contribute to those credentials (George and Cowan 1999). It may be carried out in real-time (Kowalski et al. 2009).

All such testing has as one of its functions the measurement of achievement which *ipso facto* is a measure of a specific performance. That seems to be the way it is taken by the engineering community for a crude count of papers presented at the *Frontiers in Engineering Conferences* between 2006 and 2012 in the area of assessment, yielded some 80 (of 100) that had some grammatical form of the term 'assessment' in their title that might be relevant to this text although only one them included the term 'performance'. Eleven included the term 'competence' which is clearly related to performance. Of these ten were of European origin.

Miller (1990) a distinguished medical educator said competence was the measurement of an examinee's ability to use his or her knowledge, this to include such things as the acquisition of information, the analysis and interpretation of data, and the management of patient problems (Mast and Davis 1994). Therefore, this discussion could clearly relate to the testing of academic achievement, however, in this case that is only relevant in so far as it predicts the performance of individuals in the practice of engineering, and many attempts have been made to do this (e.g. Carter 1992; Nghe et al. 2007; Schalk et al. 2011). Much of the recent concern with the assessment of performance has been driven by politicians seeking accountability in order to try and demonstrate that value is added to the students and the economy by the education they have received (Carter and Heywood 1992), and that is nothing new.

Given that the purpose of professional study is preparation to undertake work in a specific occupation a major question is the extent that a curriculum prepares or does not prepare students' to function satisfactorily in that occupation. Assessments that make this judgment may be made at both the level of the program and the level of student learning. In recent years in the United States in engineering education, there has been much focus on the program level for the purpose of accreditation. There is no equivalent to the 'clinic' used for the assessment of persons in the health professions and teaching for engineering students except in cooperative courses ('sandwich' in the UK) where the students interweave, say six months of academic study with six months of industrial training throughout a period of four years. In that case given that they should be given jobs that are appropriate to the professional function their professional performance can be assessed. This is what this writer has always taken performance-based assessment to mean, a view that has some similarity with that of Norman (1985) who said that competence describes what a person is capable of doing whereas performance is what a person does in practice. In industrial terminology, it is the equivalent of performance appraisal where like medicine and teaching there is a long history of research and practice (e.g. Dale and Iles 1992). Not that the hard data of management give one any confidence in its assessment for as Mintzberg (2009) has written "how to manage it when you can't rely on measuring it? […] is one of the inescapable conundrums of management" (p 159 and chapter). And that is the conundrum of education (Gage 1981).

Engineering educators were not largely bothered by these issues. Engineering was conceived to be the application of science to the solution of practical problems although it was not assessed as such. Educators sought answers to problems giving a single solution. It is increasingly recognized that assessments should derive from solutions to wicked problems. It is also understood that laboratory and project work contribute to the skills required by industry; the literature redounds with attempts to better relate them to practice (Heywood 2016), although Trevelyan's (2014) studies suggest there is a long way to go. There is substantial interest in problem-based learning which has been well researched (Woods et al. 1997). This chapter is focused on engineering practice and its assessment where there is a need for studies of predictive and construct validity.

Until this decade, engineering educators have paid very little attention to professional practice although this is now being rectified (Williams et al. 2013). Throughout the second half of the twentieth century there has been a strong presumption among academics that university examinations predict subsequent behaviour. Equally at regular intervals industrialists and the organizations that represent industry have complained that graduates, not just engineering graduates but all graduates are unprepared for work (performance) in industry. To put it in another way—the examinations on offer did not have predictive validity for work. What is remarkable, is the persistence of these claims over a very long period of time (fifty years), in spite of a lack of knowledge by all the partners of what it is that engineers actually 'do'. Nevertheless, from time to time during the past sixty years, attempts to find out what engineers do either directly or indirectly have been made. But first, can we learn from the past ideas, philosophies if you will, that can enable us to judge where we are in the present? It is the contention of this study that we can.

5.2 The Organization of Professional Work. Temperament and Performance

Studies of the impact of organizational structure on innovation and performance showed that often qualified engineers had poor communication skills. Engineers were required to speak several different "languages" i.e., of design, of marketing, of production, of people) (Burns and Stalker 1961). Organizational structures were shown to influence attitudes and values, and by implication performance to the extent of modifying competence although it was not perceived as such in a study by (Barnes 1960). It was posited that professional work organised in relatively open systems was likely to be more productive (innovative) than when organised in relatively closed (hierarchical systems). The findings continue to have implications for the preparation of engineers for management and leadership roles.

Immediately after the Second World War, university education in the UK was considered suitable for training engineers and technologists for R and D, but it was held that education and training for industry commonly provided by technical colleges needed to be enhanced by additional engineering science and maths, but that the art of engineering developed in industry should be retained and improved (Percy 1945). Apart from additions to content this would be achieved by sandwich (cooperative) courses developed to provide an alternative degree programme to those offered by the universities. For this purpose nine Colleges of Advanced Technology that were established to offer courses for an independently examined Diploma in Technology that would be equivalent to a university degree (Heywood 2014). Assessment was mostly an afterthought although there were one or two serious attempts to devise assessment strategies for the period of industrial training (Rakowski 1990). In practice there was a 'curriculum drift', that is to the creation of syllabuses in the mirror image of those offered by universities. Criticisms were made of these programmes by some distinguished industrialists (Bosworth 1963, 1966) that led to the discussion of at least one alternative model for the education of engineers for manufacturing engineering that was based on recent developments in the educational sciences (Heywood et al. 1966).

Throughout the 1960s the engineering profession was concerned with criticisms that scientists and technologists lacked creativity (Gregory 1963, 1972: Whitfield 1975). There were problems about how it should be taught and how it should be assessed. Later in the US, Felder (1987) listed the kind of questions that should elicit creative behaviour.

In 1962, an analysis of the end of year written examinations (five three-hour papers) taken by first year mechanical engineering students suggested that they primarily tested one factor (Furneaux 1962). Engineering drawing and coursework tested different things. It was suggested that if examinations were to be improved there was a need for their objectives to be clarified. Furneaux also sought to answer the question- were those who were tense, excitable and highly strung more likely to perform better than those were phlegmatic, relaxed and apparently well-adjusted? It might be predicted that extraverts would not do as well as introverts, for apart from anything else introverts tend to be bookish and academic studies have as their goal the development of bookish traits. Introverts work hard to be reliable and accurate, but in the extreme they take so much time on the task that they might do badly in examinations. In contrast extraverts might do an examination quickly which may be at the expense of reliability.

Furneaux found that the groups most likely to fail university examinations were the stable-extraverts, followed by (but at some distance numerically) the neurotic-extraverts. In this study the neurotic introverts did best. From this and other data, Furneaux argued that individuals who easily enter into states of high drive are likely to obtain high neuroticism scores: it is this group that is likely to obtain high examination scores. Similarly persons who have an extraverted disposition and at the same time a low drive level will do badly in examinations. An introvert with high drive will be able to compensate for relatively poor intellectual qualities whereas, in contrast, good intellectual qualities may not compensate for extraversion and low drive.

The more neurotic students did badly in the engineering drawing paper. Those who were stable did better. This, he argued, was because the task was so complex that optimum drive occurred at a low level. Furneaux found that the most common cause of failure was poor quality drawing which was of a kind that might be due to disturbing influences. Discussion with the examiner led him to believe that supra-optimal drive might have occurred, because there was some evidence of excessive sweating, lack of coordination and faulty judgement.

Irrespective of the fact that this was a small sample the finding that temperament may influence performance was found in other studies in the UK (Malleson 1964: Ryle 1969) and among engineering students in the US (Elton and Rose 1974). Given that that is the case it must also have a bearing on the way tasks are performed, which brings into question the predictive validity of a single defined competence. It goes someway to explaining Trevelyan's (2010) finding that young graduates who had entered industry from their programmes preferred solitary work. Team work was seen as splitting the assignment into parts so that each one had something to do but on his own. Anything that required collaboration was seen as an interruption. Trevelyan reported that they did not see any evidence of collaboration. He argued that the structure of the education they had received contributed to these attitudes just as the continued reception of PowerPoint lectures diminishes listening skills. But could this desire to be alone, be a function of personality? For example, do those who want to work alone have a tendency towards introversion? Yet effective engineering practice as Tilli and Trevelyan (2008) have shown requires high order skill in 'technical coordination', which requires collaboration which in turn requires a high level of interpersonal skill: other studies of practicing engineers confirm this view. Moreover, the finding that communication is a significant factor in the work of the engineers is consistent over time. But the teaching of communicative skills to engineering students was in Trevelyan's experience not done in terms of social interaction.

Development of skill in social interaction requires a high level of self-awareness and therein is the value of the peer and self-assessment of a person's performance in interpersonal activities. Early attempts to get students to self-assess in engineering focused on the ability to evaluate work they had done in completing a project.

5.3 A Multiple Strategy Approach to Assessment in Engineering Science and Its Implications for Today

In 1969 partly in response to Furneaux's criticisms of examinations in engineering, and partly in response to the ideas contained "*The Taxonomy of Educational Objectives*" (Bloom et al. 1956) a public (high school) examination in Engineering Science equivalent to at least first year university work in four year programmes was designed to achieve multiple objectives thought to represent key activities in the work of an engineer (Carter et al. 1986). The examiners had the prior knowledge that enabled them to design a new approach to assessment and the curriculum. Its underlying philosophy was that the ways of thinking of engineers were different to the ways of thinking of scientists even though engineers required a training that was substantially science based (Edels 1968). It was probably the first course in the UK to state specifically what attitudes and interests a course in engineering science should promote. The examiners believed that while not all the attitudes stated could be directly measured, it was possible to detect them in the way students tackled problems based both on the syllabus content and on coursework. The designers persuaded of the power of assessment to influence that philosophy and the changes they wished to bring about attempted to design examinations for the written papers that took into account the restraints within which engineers function in real-life situations, and they radically changed the approach to coursework and its assessment. It was found that the initial requirements overloaded the curriculum and overburdened teachers with assessment, and in consequence had to be reduced.

Eventually students were required to undertake laboratory work related to content, and keep a record of it in a journal. They had to select two experimental (discovery) investigations on different topics for assessment, and complete a 50 laboratory hour project. Rubrics were designed each of which assessed six key ability (competency) domains relevant to the investigations and separately to the project. These were preceded by six mastery assessments as appropriate to the investigations or the project. The first rubric was strictly criterion referenced (dichotomous scales). It was found that both students and teachers had difficulties with some of the items because there were no shades of grey. The key ability domains were then scaled (semi-criterion referenced) in a way that is now familiar.

The examination and coursework assessment procedure exemplify a "balanced" system of assessment described by Pellegrino et al. (2001), Heywood et al. (2007).

Dichotomous scales are not suitable for deriving grades that could be incorporated into a norm-referenced system of scoring as some critics of the scheme had envisaged. Neither are they necessarily valid. The immediate effect of moving to the new system in the year that followed was to elevate the distribution at the lower end of scale and so recognize some competence on the part of the weaker candidates. Over a fifteen year period the late D.T. Kelly, found that the coursework component discriminated well between the candidates, and was reasonably consistent from year to year (unpublished documents and Carter et al. 1986).

An analysis of some 100 papers on assessment published in the FIE Proceedings between 2006 and 2013 and in ASEE proceedings between 2009 and 2013 revealed only one study of the validity of competency statements. It seems that there is a temptation to assume that stated outcomes are valid. However, Squires and Cloutier (2011) found similarly to the engineering scientists that when the perceptions of instructors and students of the competencies addressed in a web-based campus courses in systems engineering programme were compared there were considerable discrepancies between them.

That examinations and assessment do not always test what it is thought they should test is also illustrated by the engineering science development. In order to test the skills involved in "planning" the written examination incorporated a sub-test which described a problem for which the students were required to present a plan for its solution. It was expected that this would cause them to repeat the skills they had learnt while planning their projects. There would, therefore, be a high correlation between the two marks. Unfortunately the Board's research unit did not provide a correlation between marks for project planning and evaluation in the project exercise and those for the exam that was supposed to model these skills.

But the assessors were given the correlations between coursework and the other components of the examination and the lowest correlation was found to be between these two activities. Each of three repetitions produced the same pattern of results, and the factorial analyses suggested that the sub-tests were measuring different things as was hoped (Heywood and Kelly 1973). Similar results were obtained in each of three consecutive years. At first it was thought that this effect was because engineering design was not a requirement of the syllabus. However, a decade later when Sternberg published his Triarchic theory of intelligence another explanation became possible (Sternberg 1985). Sternberg distinguished between three components of intelligence: *meta-components* which are processes used in planning, monitoring and decision-making in task performance; *performance components* which are processes used in the execution of a task; and *knowledge-acquisition components* used in learning new information. Each of these components is characterized by three properties—duration, difficulty and probability of execution which are in principle independent.

It is evident that the project assessment schemes are concerned with the evaluation of meta-components. Elsewhere he calls them appropriately "*executive processes*". We can see that a key difference between the project planning exercises and the written sub-test is the time element. The two situations required the student to use different information processing techniques. The written exercise is a different and new domain of learning for which training is required. In order for the skill to become an old domain, a high level of automatization is required so that the different processes in the meta-component are brought into play more quickly. That is to say they have at a certain level to become non-executive. The project and the written paper while demanding the same meta-components might be regarded as being at different levels in the experiential learning continuum. Some executive processing will always be required at the written paper level, and it is possible to argue that the task performance and stress which it creates are a more accurate reflection of the everyday activities of executives than the substantive project.

Notwithstanding the validity of this interpretation, more generally these investigations showed that when criterion and semi-criterion measures appear to have high face validity there is a need to ensure that there is congruence between student and assessor perceptions of the items. The measurement of performance is not as easy as it seems. Similarly it cannot be assumed that the goals that assessors have for multiple-strategy assessments are necessarily being met even when they also appear to have face validity.

5.4 The Development of Competency-Based Qualifications in the UK and the US in the 1980s and 1990s

Throughout the 1980s and 1990s in the UK and the US there were developments in the provision of competency-based qualifications. In the UK National Vocational Qualifications were introduced. An NVQ "is a statement of competence clearly relevant to work and intended to facilitate entry into, or progress in, employment and further learning, issued to an individual by a recognized awarding body" (Jessup 1991, p. 15).

"The two aspects of the statement of competence which the definition (above) goes on to add are significant, as is their relationship to each other. The statement leads on "performance", which is of course central to the concept of competence, and places 'skills, knowledge and understanding' as underpinning requirements of such performance. This does not deny the need for knowledge and understanding, but does make clear that they, however necessary, are not the same thing as competence. This position has considerable implications for assessment" (Jessup, p. 16). The assessors of engineering science paid great attention to the student's theoretical understanding and some teachers would have said too much. This remains a critical issue especially when successful projects seem to be based on inadequate understanding of the theoretical rationale!

Jessup writes: "Assessment may be regarded as the process of collecting evidence and making judgements on whether performance criteria have demonstrated he or she can meet performance criteria for each element of the competence specified" (Jessup. p. 18).

Jessup's primary thesis was to argue that the outcomes-based approach that he and others had developed was applicable to all forms of learning. The system that emerged had five levels of competence and it may be argued that the Bologna levels have their origins in this kind of thinking. Among the techniques of assessment discussed at the time were records of achievement which in the past couple of years have been introduced in the UK Universities (HEAR 2014).

Tyler's and Bloom's approaches were rejected because they chose outcomes that could easily be assessed. However, engineering educators like medical educators are unlikely to assume that only those things that are defined are the only elements of competence. During this period a report from the Department of Employment included a taxonomy of outcomes for engineering education by Carter (1984, 1985) that embraced the cognitive and affective domains (Otter 1992). Attention is drawn to a *Taxonomy of Engineering Education Outcomes* that embraces both the cognitive and affective domains.

An alternative to the competency approach was advocated by the Royal Society for Arts and Manufacture (RSA) under the title "Education for Capability". The basis of the project was action learning as a means of helping students learn how to apply their knowledge and skills. It argued that students should be able to negotiate their programmes of study, should learn through collaborative learning, and learn skill in reflection through reflection on their progress. Capability may be assessed by observing if students are able to take effective action, explain what they are doing, live and work effectively with others, show that they learn from experience be it their own or with others (Stephenson and Weil 1992: Stephenson and Yorke 1998). In contrast to the competency approach capability is holistic and broader concept than competency.

In the US, there was a movement to develop 'standards' in schools. In essence they were very long lists of outcomes. It was argued that helpful though they may be they were too long for teachers to contemplate covering. In 2000, The International Technology Education Association published a list of standards for technological Literacy. There has been interest in recent years in producing a corresponding list for engineering.

In 1989, the UK Employment Department introduced the Enterprise in Higher Education Initiative in universities with the intention that all departments in a university would arrange their curriculum so that it would develop the "personal transferable skills that were considered to be required by industry. They did not think it was necessary to add bolt on subjects, and a philosophy of assessment was suggested" (Heywood 2005).

In 1992, a report to the US Secretary of Labor argued that the high school curriculum was not equipping students for the world of work (SCANS 1992). To achieve this goal, students required to develop five work place competencies: handling of resources; interpersonal skills; handling information; thinking in terms of systems, and technology. The SCANS Committee believed that their goals could be achieved by adjusting teaching within the ordinary subjects of the curriculum and gave examples of how this might be accomplished. A weakness of the model was that it paid little attention to the "practical" areas of the curriculum like the arts and crafts and music.

The American College Testing program (ACT) was involved in developing tests for SCANS and the program attracted the interest of some engineering educators (Christiano and Ramirez 1993). Much earlier, ACT developed a College Outcome Measures Program (COMP) (ACT 1981; Forrest and Steele 1982) which was designed as a measure of general education. It was designed to help institutions evaluate their curricular and/or design learning activities which would help students obtain the knowledge skills and attitudes necessary for functioning in adult roles after graduation. These aims go beyond education for work and take into account more general aspects of living (Lotkowski et al. 2004).

Taken together there are remarkable similarities between the objectives of these different programmes and the concept of intelligence as derived from the views of experts and lay people by Sternberg (1985).

5.5 Studying Engineers at Work

Attempts to develop new curricular in engineering were and are criticized because they are based on models of what it is believed engineers do rather than actual studies of what they do. An early analysis of the work done by engineers a highly innovative firm in the aircraft industry in the UK had as its prime purpose the derivation of a general method for the determination of the objectives of training technologists and technicians (Youngman et al. 1978). A secondary purpose took into account factors such as satisfactory performance on the job, and organizational structure in order to show how work could be structured for training purposes. Fourteen engineering activities and twelve work types were identified. The work types did not match with textbook models of the engineering process. Significantly, no traditional manager work type emerged. One interpretation of the data argued that to some extent everyone in an engineering function engaged at one level or another in management. It was found that that age and job level were more significant variables than educational qualifications in terms of explaining differences in job descriptions. This analysis tended to support this view that as engineers grow older they tend to place increasing reliance on experience and reject the notion that training can be beneficial. It was suggested that over reliance on experience could impede innovation. A view of the firm as a learning organization was described.

The study did not result in a taxonomy but *The Taxonomy of Educational Objectives* was shown to be of value in the analysis of tasks done managers in a steel plant (W. Humble cited by Heywood 1970). It showed the importance of the affective domain. A survey by the Engineering Industries Training Board showed the importance of this domain since it found that 60 % of technologists time were spent in interpersonal activities thus confirming the importance of interpersonal competence. This finding was also repeated in a comparative study of German and British production managers (Hutton and Lawrence 1981). The similarities between the two groups were found to be much greater than the differences. A problem for the German engineers was dealing with critical incidents. A problem for British engineers was resources. The Germans tended to emphasize the technical whereas the British emphasized the managerial aspects of the job. In both cases the paradox was of jobs that were fragmented during the day but nevertheless coherent. In contrast to the earlier study where a relationship between status and morale had been found no such relationship was found in the British firms studied.

The findings of these studies support the work of more recent authors like Bucciarelli (1994), and Vincenti (1990) and in particular the studies reported in Williams et al. (2013) to the effect that engineering is far more than the application of science and a messy activity when compared with the search for truth. It is a social activity and because of that interpersonal competence is a skill to be highly valued.

5.6 Intellectual, Emotional and Professional Development

It is therefore of some consequence that the curriculum has tended to neglect the affective dimension of behaviour at the expense of the cognitive. During the 1980s some engineering educators engaged in discussions about these dimensions. Efforts were made to design curricula that would respond to Perry's model of intellectual development (Culver and Hackos 1983; Marra et al. 2000; Pavelich and Moore 1996). A curriculum designed to follow the stages of this model should lead students from dependence to independence where they take responsibility for their own learning, and are able to solve ambiguous problems such as are posed by real-life engineering. To achieve this competence, it is argued that students need to be reflective practitioners but Cowan counsels that Schön from whom the idea of reflective practice comes, does not take account of "reflection-for-action." (Cowan 2006). Cowan distinguishes between nine levels of reflection. Significantly he notes that reflection requires the ability to be self-aware and its key skill is to be able to self-assess, and it is this that distinguishes it from analysis which occupies so much of engineering education.

How engineering subjects are taught matters for the development of professional practice for such practice depends on judgment and judgement needs to be reflective. Work is both cognitively and emotionally construed for which reason it is incumbent on employers and employees to understand how organizations and the individual interact at this level, and they could do no better than look at the 1960 reports from Barnes (1960) and Burns and Stalker (1961). Some engineering educators have discussed critical thinking within the context of reflective practice (Mina et al. 2003). It is safe to conclude that reflective practice and critical thinking are best developed and assessed when they are built into the whole curriculum.

No discussion of the affective domain can ignore the writings on emotional intelligence. Whatever you may think about it as unitary concept it is clear that we have to handle its components every day (Culver 1998; Goleman 1994; Bar-On and Parker 2000). 'Its' development can be assisted in both education and training but it cannot be left to education alone because education cannot simulate the everyday situations that have to be faced in industry and its learning is part of the process of development.

Intellectual, emotional and professional development cannot be completed within college alone for a person continues to develop and will do so in response to any situation. For this reason industrialists have as much responsibility for the development of their engineers as do the colleges from which they come, and in these days of rapid turnover have an obligation to help them prepare for their next assignment.

5.7 Other Aspects of Outcomes Assessment

By 2000, the engineering curriculum had come to be and continues to be dominated by outcome approaches to assessment. *The Taxonomy of Educational Objectives* which despite criticisms of its use in engineering continues to influence some educators but its 2001 revision is beginning to be of interest (Anderson et al. 2001). Its use in computer science has been questioned (Highley and Edlin 2009). But there have been attempts to analyze questions set in examinations to evaluate the extent to which critical thinking and problems solving skills were being tested (Jones et al. 2009). But the judgments were based on face validity. That engineers read beyond the subject is demonstrated by a paper that describes the design of a service course using Fink's (2003) taxonomy which at sight shows as much concern for the affective domain as it does for the cognitive (Fero 2011).

An attempt to reflect the variability of student performance that indicates what objectives students should attain at three different levels of performance has been reported by Slamovich and Bowman (2009). Outcomes assessment surveys have come to be used as a means of programme evaluation. An unusual one reported on the use of the MBTI (personality indicator) to assign students to teams (Barr et al. 2006). A substantial study of entrepreneurially minded engineers that embraced the affective domain compared practicing engineers with students. The practicing engineers demonstrated a different 'footprint'. They had lower interpersonal skills, lower creativity, lower goal orientation and lower negotiating skills (Pistrui et al. 2012).

Prior to ABET EC 2000, a study at Iowa State University sought to find out how to assess ability (Brumm et al. 2006). Workplace competencies were defined as the application of knowledge, skills, attitudes and values and behaviour that interact with the task at hand, 'Key actions' required to demonstrate a competency were described. Among other findings, the authors suggest that from a combination of supervisor and self-assessments an e-career self-management system could be developed. While the study was based on experiential learning there is no mention of a taxonomy in this area. Neither the taxonomy presented nor the Iowa study makes any mention of the skill of 'reflection'.

A European (University of Vienna) comparison of competencies valued by employers of computer science graduates and faculty showed that whereas teachers highly valued analysis and synthesis, basic general knowledge and research skills, employers valued capacity to learn, team competence, ability to work autonomously, problem solving, and interpersonal skills and ability to work with new technology. (Kabicher and Motschnig-Pitrik 2009). This raises questions of research design for it might be thought that the acquisition of research skills would necessarily involve problem solving.

European literature shows that irrespective of the term used an outcomes approach will generate lists that are common across the globe. One Spanish study reduced 37 core competencies to 5 (Chadha and Nicholls 2006—see also Tovar and Soto 2010). The three Viennese studies recognized the importance of the affective domain in the development of competencies in computer science. In one study, a technique for getting students to share their reflections with each other is described. Students in another study reported how the skills learnt had benefited their private lives as well as their professional (Motschnig-Pitrik and Figl 2007). It was found that it was not possible to provide for team projects and assume that all that could be learnt from them was learnt (Figl and Motschnig-Pitrik 2008). Distinctions were made between specific and generic task team competencies, and knowledge, attitude and skills competencies. These studies lead to the view that students could benefit from courses in learning-how-to-learn: Motschnig-Pitrik and Figl (2007) suggest that a course in 'soft skills' lead to an enhancement of these competencies as perceived by the students when functioning in teams.

One of the important findings of an evaluation of a competency-specific engineering portfolio studio in which students selected the competencies they wished to develop was that pedagogies that support individual choice involve a shift in the power dynamics of the classroom (Turns and Sattler 2012).

Related to the concept of the hidden curriculum is the idea of "accidental competencies" promoted by Walther and Radcliffe (2006). They are "attributes not achieved through targeted instruction". It is a concept that has similarities with Eisners concept of "expressive outcomes".

5.8 Project Work and Teamwork

Project work and teamwork are considered to be effective ways of developing the 'soft' skills (now called professional skills in the US) that industry requires. In addition carefully selected projects can help students work across the disciplines. It has been found that high levels of interdisciplinarity and integration may contribute to positive learning experiences (Froyd and Ohland 2005). However, it is suggested that many students are challenged by collaboration skills. Such skills have a large affective component and are context dependent as a function of an individual's personality. Communication skills are particularly challenged when groups have different perceptions of the problem. Transdisciplinary projects are able to integrate the tools, techniques and methods from a variety of disciplines. Impediments to collaboration include disciplinary prejudice, unwillingness to listen and ask questions and lack of shared ideas (Lingard and Barkataki 2011). A key problem that is not fully understood is the level of knowledge required by each partner in the other disciplines involved. "Constructive Controversy" has been recommended as means of creating mutual understanding about a problem area (Johnson and Johnson 2007; Matusovich and Smith 2009). An experimental course based on constructive controversy led to the reminder that the pedagogic reasoning for the use of non-traditional methods of instruction needs to be explained (Daniels and Cajander 2010).

It has been argued that teamwork can contribute to the development of innovation skills and creativity. The research reported on the former suggested that heterogeneous teams were not more innovative than homogenous teams (Fila et al. 2011). One study that used design notebooks found that contrary to experience the most efficient use of the creative process was in the production phase and not in the conceptual design phase. In both areas of innovation and creativity there needs to be more research (Ekwaro-Osire et al. 2009). Importance has been attached to self and peer assessment. One study reported that students only made a fair job of self-assessment and this did not change with time. A voluntary system of self-assessment was used to support a variety of learning styles. Those who used the assessment system performed better than those who had not (Kemppaineme and Hein 2008). Another study found that whilst students could distinguish between good, average and poor performance they could not transfer their views to marking on a standards scale. A great deal of care has to be taken in areas where assessors are likely to disagree, as for example with oral presentations (Wigal 2007). There is increasing use of on-line peer review schedules, and one study reports writing instruction could be improved with the introduction of a peer-review component (Delson 2012).

The final section asks whether or not teamwork can be taught? Since participation in teamwork requires the utilisation of a number of skills that can be practiced the answer is yes. Similarly if students have some knowledge of learning they may also be helped to better their participation in team activities. Team activities should be seen as a preparation for industry in the sense that it should enable the person to better understand and evaluate the situation in which he or she finds himself or herself.

5.9 Preparation for Work

The belief that competence is a trait that individuals possess to one degree or another prevails in engineering education. It leads to the view that students can be prepared for work immediately on graduation by the acquisition of specifically stated competencies that can be taught. This belief been challenged on several occasions. For example a phenomenological study of engineers of work is reported by Sandberg (2000) that offers an alternative view of competency furthers this view. Competency is found to be context dependent and a function of the meaning that work has for the individual involved. Engineers were found to have different perceptions of work, and competencies related to the same task were found to be hierarchically ordered among them, each level being more comprehensive than the previous level. Attributes are developed as a function of work. It follows that they are not fixed, therefore firms will have to undertake training (or professional development) beginning with an understanding of the conception that the engineers has of her/his work. Professional competence should be regarded as reflection in action or understanding of work or practice rather than as a body of scientific knowledge.

Little has been known about how engineers utilize the knowledge learnt in their educational programmes at work. A study is reported by Kaplan and Vinck (2013) that affirms previous findings that engineers tend to use off-the-shelf solutions or start with an analogy of an existing solution for a different problem. The same authors noted that engineers switch between scientific and design modes of thinking. Yet another study reported the view that engineers who are contextually competent are better prepared for work in a diverse team.

These studies revealed some important competencies that are rarely discussed. For example, the "ability to see another person's point of view". This requires that the teaching of communication skills should be done in terms of social interaction. The key skill of "technical coordination" found by Trevelyan depends equally on the possession of communication and interpersonal skills.

Surprisingly studies of the use of mathematics by engineers are also shown to be related to the affective domain and influenced by sociocultural forces (Gould and Devitt 2013). Tacit knowledge is found to be important.

5.10 Conclusion

There are three groups of professionals with whom engineers can be compared—management, medicine and teaching. They differ from engineers in that in order to receive professional recognition they have to demonstrate they can perform the tasks required of them to a satisfactory standard. In the case of doctors and teachers, they receive substantial clinical training during their degree programme. There is no equivalent to this in engineering programmes unless they are structured as cooperative courses. Across the world the regulators of the engineering curriculum require an outcomes/competency-based approach to assessment (ABET; Bologna; Engineers Australia; Engineers Ireland; Engineering Council; Tuning).

Studies of engineers at work show the complexity of the activity of engineering. The idea that an engineer should be able to take on a professional role immediately on leaving college without some prior guided experience of working in industry is shown to be nonsense. It supports the view that engineering education may better prepare students for work in industry if it is structured on a cooperative basis. But the experience of industry has to be carefully designed. Alternative support for this view is to be found in Blandin's (2011) study of a cooperative course that upgraded technicians to technologist status. He found that within the company the students developed competencies that were specific to their job. This writer takes from this study that the interaction between periods of academic study and industrial work help students to acquire professional competence in professional engineering that is not available to courses of the traditional kind that have no industrial contact.

It seems clear that the acquisition of a competence is through a developmental process and subject to the experience of the conditions imposed on it by the organisation. This has implications for way formative assessment is conducted and the way the curriculum is designed.

Since this chapter was written, ABET has made proposals for changing its requirements. These have not been greeted with aplomb by many engineering educators with the effect that a debate is now in progress. At the same time the literature on assessment continues to amass (Heywood 2016).

Issues/Question for Reflection

- The interaction between periods of academic study and industrial work help students to acquire professional competence in professional engineering
- Team activities should be seen as a preparation for industry in that they should enable the person to better understand and evaluate a given situation
- Should engineering students be given courses in learning and assessment especially self-assessment?
- What is the best way to assess competence in engineering?
- Do industrialists have as much responsibility for the development of their engineers as do the colleges from which they come?

References

ABET (Accreditation Board for Engineering and Technology) (1997) EC 2000. *Engineering Criteria*.

ABET. (2010). 2010 Annual ABET report. www.abet.org

ACT. (1981). *Defining and measuring general education, knowledge and skills. COMP technical report 1976–1981*. Iowa, City: American College Testing Program.

Anderson, L. W., Krathwohl, D. R., et al. (Eds.). (2001). *A Taxonomy for learning teaching and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley/Longman.

Anderson, L. W., & Sosniak, L. A. (Eds.). (1994). *Bloom's taxonomy a forty-year retrospective*. Chicago: National Society for the Study of Education. University Press of Chicago.

Barnes, L. B. (1960). *Organizational systems and engineering groups. A comparative study of two technical groups in industry*. Boston: Harvard University, Graduate School of Business Administration.

Barr, R. E, Krueger, T. J., & Aanstoos, T. A. (2006). Continuous outcomes assessment in an introduction to mechanical engineering course. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. S1E-9 to 14.

Bar-On, R., & Parker, D. A. (Eds.). (2000). *The handbook of emotional intelligence*. San Fransisco: Jossey Bass.

Besterfield-Sacre, M., et al. (2000). Defining outcomes. A framework for EC-2000. *IEEE Transactions in Education, 43*, 100–110.

Blandin, B. (2011). The competence of an engineer and how it is built through an apprenticeship program: A tentative model. *International Journal of Engineering Education, 28*(1), 57–71.

Bloom, B. S. et al. (Eds.). (1956). *The Taxonomy of Educational Objectives. Handbook 1. Cognitive Domain*. New York: David Mackay. (1964). London: Longmans Green.

Bologna. (1999). The Bolgna Process. Towards the European Higher Education Area. http://cc.europa.cu/education/policies/educ/bologna/bolgna_en.html

Bologna. (2005). Working group on qualifications frameworks. A framework for qualifications of the European Higher Education Area. http://www.bologna-bergen2005.no/Docs/00-main_doc/050218_QF_EHEA.pdf

Bosworth, G. S. (1963). Toward creative activity in engineering. *Universities Quarterly, 17*, 286.

Bosworth, G. S. (Chairman of a Committee) (1966). *The Education and Training requirements for the Electrical and Mechanical Manufacturing Industries. Committee on Manpower Resources for Science and Technology*. London: HMSO.

Brumm, T. J., Hanneman, L. F., & Mickelson, S. F. (2006). Assessing and developing program outcomes through workplace competencies. *International Journal of Engineering Education, 22*(1), 123–129.

Bucciarelli, L. L. (1994). *Designing Engineers*. Cambridge MA: MIT Press.

Bucciarelli, L. L., Coyle, C., & McGrath, D. (2009). Engineering education in the US and EU. In S. H. Christensen, B. Delahousse, & M. Meganck (Eds.), *Engineering in context*. Academica: Aarhus, Denmark.

Burns, T., & Stalker, G. (1961). *The Management of innovation*. London: Tavistock.

Carberry, A. R., Siniawski, M. T., & Dionisio, D. N. (2012). Standards based grading. Preliminary studies to quantify changes in affective and cognitive student behaviour. In *Proceedings Frontiers in Education Conference*, ASEEE/IEEE. pp. 947–951.

Carter, G. (1992). The fall and rise of university engineering education in England in the 1980s. *International Journal of Technology and Design Education, 2*(3), 2–21.

Carter, G., & Heywood, J. (1992). The value-added performance of electrical engineering students in a British university. *International Journal of Technology and Design Education, 2*(1), 4–15.

Carter, G., Heywood, J., & Kelly, D.T. (1986). *A Case Study in Curriculum Assessment. GCE Engineering Science (Advanced)*. Manchester: Roundthorn Publishing.

Carter, R. G. (1984). Engineering curriculum design. *Institution of Electrical Engineers Proceedings*, 131, Part A p 678.

Carter, R. G. (1985). Taxonomy of objectives for professional education. *Studies in Higher Education* 10(2).

Chadha, D., & Nicholls, G. (2006). Teaching transferable skills to undergraduate engineering students. Recognizing the value of embedded and bolt-on approaches. *International Journal of Engineering Education, 22*(1), 116–122.

Christiano, S. J. E., & Ramirez, M. (1993). Creativity in the classroom. Special concerns and insights. *Proceedings Frontiers in Education Conference,* ASEE/IEEE. pp. 209–212.

Cotgrove, S. (1958). *Technical education and social change*. London: Allen and Unwin.

Cowan, J. (2006). *On becoming an innovative university teacher* (2nd ed.). Buckingham: Open University Press.

Cox, R. (1967). Examinations and higher education. A review of the literature. *Universities Quarterly*, 21(3), 292.

Culver, R. S. (1998). A review of emotional intelligence by Daniel Golman. Implications for technical education. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. pp. 855–860.

Culver, R. S., & Hackos, J. T. (1983). Perry's model of intellectual development. *Engineering Education, 73*(2), 221–226.

CVCP. (1969). *The Assessment of academic performance*. London: Committee of Vice-Chancellors and Principals.

Dale, M and P. Iles (1992). *Assessing management skills. A guide to competencies and evaluation techniques*. London, Kogan Page.

Daniel, W. W., & McIntosh, N. (1972). *The right to manage*. London: Macdonald and James.

Daniels, M., & Cajander, A (2010). Experiences from using constructive controversy in an open ended group project. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. S3D-1 to 5.

Delson, N. (2012). RATEMYTEAMMATE.Org: A proposal for an on-line tool for team building and assessment. In *Proceedings Annual Conference American Society for Engineering Education.* Paper 5024.

Edels, H. (1968). *Technology in the sixth-form. Trends in Education*. No 10. April. London: Ministry of Education.

Ekwaro-Osire, S., Mendias, J. J., & Orono, P. (2009). Using design notebooks to map creativity during team activities. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. M1J-1 to 5.

Elton, C. F., & Rose, H. A. (1974). Students who leave engineering. *Engineering Education, 62* (1), 30–32.

Engineers Australia. (2009). *Australian Engineering Competency Standards*. http:www.engineersAustralia.org.au/shadomx/apps/fms/fmsdownload.cfm?file_uuid=600A19c1-FE15-BEB8-BC7B-0941CCOF1020&siteName=iaust

Engineers Ireland. *Regulations for the title Chartered Engineer*. Dublin: Engineers Ireland.

Engineering Council. *UK Standards for Professional Engineering Competence*. London: Engineering Council.

Felder, R. M. (1987). On creating creative engineers. *Engineering Education, 74*(4), 222–227.

Ferro, P. (2011). Use of Fink's Taxonomy in establishing course objectives for a re-designed materials engineering course. In *Proceedings Annual Conference of the American Society for Engineering Education*, Paper 309.

Figl, K., & Motschnig-Pintrik, R. (2008). Researching the development of team competencies in computer science courses. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. S3F-1 to 6.

Fila, N. D., Wertz, R. E., & Purzer, S. (2011). Does diversity in novice teams lead to greater innovation? In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. S3H-1 to 5.

Fink, L. D. (2003). *Creating significant learning experiences. An Integrated approach to designing college courses*. San Fransisco: Jossey Bass.

Forrest, A., & Steele, J. M. (1982). *Defining and measuring general education knowledge and skills. Technical Report*. American College Testing Program (ACT): Iowa City.

Furneaux, W. D. (1962). The psychologist and the university. *Universities Quarterly, 17*, 33–47.

Froyd, J. E., & Ohland, M. W. (2005). Integrated engineering curricula. *Journal of Engineering Education, 94*(1), 147–164.

Gage, N. L. (1981). *Hard gains in soft science. The case of pedagogy*. Bloomington, Ind: Phi Delta Kappa.

George, J., & Cowan, J. (1999). *A handbook of techniques for formative evaluation: mapping the students learning experience*. London: Kogan Page.

Goleman, D. (1994). *Emotional intelligence. Why it matters more than IQ*. New York: Bantam Books.

Gould, E., & Devitt, F. (2013). Mathematics in engineering practice. Tacit trumps tangible. In Williams, Figueiredo and Trevelyan (Eds.). *Engineering practice in a global context. Understanding the technical and the social*. London. CRC Press/Taylor and Francis.

Gregory, S. A. (1963). Creativity in chemical engineering research. *Transactions of Institution of Chemical Engineers*, Paper 650, p 70.

Gregory, S. A. (Ed.). (1972). *Creativity and innovation in engineering*. London: Butterworths.

Hamilton, A. (2012). This postgraduate brain drain needs plugging. *The Times*, March 1st. p. 18.

Hartle, T. W. (1986). The growing interest in measuring the education achievement of college students. In C. Adelman (Ed.), *Assessment in higher education*. Washington, DC: US Department of Education.

Hayes, S. V., & Tobias, S. A. (1964/5). The project method of teaching creative mechanical engineering. In *Proceedings of the Institution of Mechanical Engineers,* p. 179.

HEAR. (2014). *Higher Education Achievement Report (HEAR)*. www.hearca.uk. Accessed 10th Aug 2014.

Hesseling, P. (1966). *A strategy for evaluation research*. Assen, Netherlands: Van Gorcum.

Heywood, J. (1970). Qualities and their assessment in the education of technologists. *International Bulletin of Mechanical Engineering Education., 9*, 15–29.

Heywood, J., Carter, G., & Kelly, D. T. (2007). Engineering science A level in the UK. A case study in the balanced assessment of student learning, educational policies and educational scholarship. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. S4f–9 to 13.

Heywood, J., & Kelly, D. T. (1973). The evaluation of course work- a study of engineering science among schools in England and Wales. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. pp. 269–276.

Heywood, J. (2005). *Engineering education. Research and development in curriculum and instruction*. Hoboken, NJ: IEEE/Wiley.

Heywood, J. (2014). Higher technological education and British policy making. A lost opportunity for curriculum change in engineering education. In *Proceedings Annual Conference American Society for Engineering Education*. Paper 8689.

Heywood, J. (2016). *The assessment of learning in engineering education. Practice and policy*. Hoboken, NJ: Wiley/IEEE Press.

Heywood, J., et al. (1966). The education of professional engineers for design and manufacture (A model curriculum). *Lancaster Studies in Higher Education, 1*, 2–151.

Highley, T., & Edlin, A. E. (2009). Discrete mathematics assessment using learning objectives based on Bloom's taxonomy. In *Proceedings Frontiers in Education Conference,* M2J–1 to 6.

Hutton, S. P., & Lawrence, P. A. (1981). *German engineers. The anatomy of a profession*. Oxford: Clarendon Press.

Jessup, G. (1991). *Outcomes. NVQ's and the emerging model of education and training*. London: Falmer.

Johnson, D., & Johnson, R. (2007). *Creative constructive controversy. Intellectual challenge in the classrooms* 4th Edn. Edina Min: Interaction pbl.

Jones, K. O., Harland, J., Reid, J. M. V., & Bartlett, R. (2009). Relationship between examination questions and Bloom's taxonomy. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. WIG-1 to 6.

Kabicher, S., & Motschnig-Pitrik, R. (2009). What competences do employers, staff and students expect from a computer science graduate? In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. W1E–1 to 6.

Kaplan, F., & Vinck, D. (2013). The practical confrontation of engineers with a new design endeavour. The case of the digital humanities. Ch 3 of Williams, Figueiredo and Trevelyan (Eds.). *Engineering practice in a global context. Understanding the technical and the social*. London. CRC Press/Taylor and Francis.

Kemppainem, A., & Hein, G. (2008). Enhancing student learning through self-assessment. In: *Proceedings Frontiers in Education Conference*, ASEE/IEEE. T2D-14 to 19.

Kowalski, S. E., Kowalski, F. V., & Gardner, T. Q. (2009). Lessons learned when gathering real-time formative assessment in the university classroom using Tablet PC's. In *Proceedings Frontiers in Education Conference*, T3F-1 to 5.

Lingard, R., & Barkataki, A. (2011). Teaching teamwork in engineering and computer science. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. F1|C-1 to 5.

Lotkowski, V. A., Robbins, S. B., & Noeth, R. (2004). *The role of academic and non-academic factors in improving college retention. ACT Policy Report*. ACT: Iowa.

Malleson, N (1964). *A Handbook of British student health services*. London, Pitman.

Marra, R., Palmer, B., & Litzinger, T. A. (2000). The effects of first year design course on student intellectual development as measured by the Perry Scheme. *Journal of Engineering Education, 89*(1), 39–45.

Mast, T., & Davis, D. A. (1994). Concepts of competence. In D. A. Davis & R. D. Fox (Eds.), *The physician as learner. Linking research to practice*. Chicago: American Medical Association.

Matusovich, H., & Smith, K. (2009). Constructive academic controversy-What is it? Why use it? How to Structure it. In *Proceedings Frontiers in Education Conference,* M3A–1 to 3.

McGuire, C. (1993). Perspectives in assessment. In J. S. Gonnella, et al. (Eds.), *Assessment measures in medical school residency and practice*. New York: Springer.

McVey, J. (1975). The errors in marking examination scripts in electronic engineering. *International Journal of Electrical Engineering Education, 12*(3), 203.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, 563–567.

Mina, M., Omidvar, R., & Knott, K. (2003). Learning to think critically to solve engineering problems: revisiting John Dewey's ideas for evaluating engineering education. In *Proceedings Annual Conference American Society for Engineering Education*. Paper 2132.

Mintzberg, H. (2009). *Managing*. Harlow: Pearson.

Motschnig-Pitrik, R., & Figl, K. (2007). Developing team competence as part of a person centered learning course on communication and soft skills in project management. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. F2G-15 to 21.

NAE. (2004). *The engineer of 2020. Visions of Engineering in the New Century*. Washington, DC: National Academies Press.

NCVQ. (1989). *NVQ criteria and procedures*. London: National Council for Vocational Qualifications.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. T2G-7 to 12.

Norman, G. R. (1985). Defining competence. A methodological review. In Neufeld, V. R. & Norman G. R. (Eds.) *Assessing clinical competence.* New York: Springer.

Otter, S. (1992). *Learning outcomes in higher education*. London: HMSO for the Employment Department.

Oxtoby, R. (1973). Engineers, their jobs and training needs. *The Vocational Aspect of Education, 25*, 49–59.

Pavelich, M. J., & Moore, W. S. (1996). Measuring the effect of experiential education using the Perry model. *Journal of Engineering Education, 85*, 287–292.

Pellegrino, J. W., et al. (Eds.). (2001). *Knowing what students know? The Science and design of educational assessment*. Washington, DC: National Research Council, National Academies Press.

Percy, Lord Eustace (Chairman of Committee). (1945). *Higher technological education*. London: HMSO.

Pistrui, D., Layer, J. K., & Dietrich, S. L. (2012). Mapping the behaviors, motives and professional competencies of entrepreneurially minded engineers in theory and practice: an empirical investigation. In *Proceedings Annual Conference of the American Society for Engineering Education,* paper 4615.

Rakowski, R. T. (1990). Assessment of student performance during industrial training placements. *International Journal of Technology and Design Education, 1*(3), 106–110.

Ryle, A. (1969). *Student casualties*. London: Allen Lane.

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy and Practice, 14*(3), 387–392.

Sandberg, J. (2000). Understanding human competence at work. An interpretive approach. *Academy of Management Journal, 43*(3), 9–25.

SCANS. (1992). *Learning a living. A blue print for high performance*. US Department of Labor: Washington DC.

Schalk, P. D., Wick, D. P., Turner, P. R., & Ramsdell, M. W. (2011). Predictive assessment of student performance for early strategic guidance. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. S2H-1 to 5.

Slamovich, E. B., & Bowman, K. J. (2009) All, most or some: Implementation of tiered objectives for ABET assessment in an engineering program. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. T3C-1 to 6.

Squires, A. F., & Cloutier, A. J. (2011). Comparing perceptions of competency knowledge development in systems engineering curriculum: a case study. In *Proceedings Annual Conference American Society for Engineering Education*. Paper 1162.

Stephenson, J., & Weil, S. (1992). *Quality in learning. A capability approach in higher education*. London: Kogan Page.

Stephenson, J., & Yorke, M. (1998). *Capability and quality in higher education*. London: Kogan Page.

Sternberg, R. S. (1985). *Beyond IQ. A triarchic theory of intelligence*. Cambridge: Cambridge University Press.

Tilli, S., & Trevelyan, J. P. (2008). Longitudinal study of Australian engineering graduates: preliminary results. In *Proceedings Annual Conference American Society for Engineering Education.* Paper 1537.

Trevelyan, J. P. (2010). Restructuring engineering from practice. *Engineering Studies, 2*(3), 175–195.

Trevelyan, J. P. (2014). *The making of an expert engineer*. Leiden, Netherlands: CRC Press.

Tovar, E., & Soto, O. (2010). The use of competences assessment to predict the performance of first year students. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. F3J-I to 4.

Tuning. (2005). Tuning educational structures in Europe project. Approaches to teaching, learning and assessment in competence based degree programmes. http://www.uniedusto.org/tuning

Turns, J., & Sattler, B. (2012). When students choose competencies: insights from the competence-specific engineering portfolio studio. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. pp. 646–651.

Tyler, R. W. (1949). Achievement testing and curriculum construction. In E. G. Williamson (Ed.), *Trends in student personnel work*. Minneapolis: University of Minnesota.

Vincenti, W. G. (1990). *What engineers know and how they know it. Analytical studies from aeronautical history*. Baltimore, MD: The Johns Hopkins University Press.

Walther, J., & Radcliffe, D. (2006). Engineering education: targeted learning outcomes or accidental competencies. In *Proceedings Annual Conference American Society for Engineering Education*, paper 1889.

Whitfield, P. R. (1975). *Creativity in industry*. Harmondsworth: Penguin.

Wigal, C. M. (2007). The use of peer evaluations to measure student performance and critical thinking ability. In *Proceedings Frontiers in Education Conference*, ASEE/IEEE. S3B-7 to 12.

Williams, B., Figueiredo, J., & Trevelyan, J. P. (Eds) (2013). *Engineering practice in a global context. Understanding the technical and social*. London: CRC Press/Taylor and Francis.

Woods, D. R et al (1997). Developing problem solving skills. The McMaster problem solving program. *Journal of Engineering Education, 86*(2), 75–91.

Yokomoto, C. F., & Bostwick, W. D. (1999). Modelling the process of writing measurable outcomes for EC 2000. In *Proceedings Frontiers in Education Conference,* ASEE/IEEE. 2, 11b-8–22.

Youngman, M. B., Oxtoby, R., Monk, J. D., & Heywood, J. (1978). *Analysing jobs*. Aldershot: Gower Press.

Chapter 6
Instruction and Assessment of Competencies: Two Sides of the Same Coin

Paul F. Wimmers, Lourdes R. Guerrero and Susan Baillie

Abstract Six core competencies, as defined by the Accreditation Council for Graduate Medical Education (ACGME), offer a conceptual framework to address the knowledge and skills needed by students in training and doctors to perform competently. The question of interest for educators is "how" residents perceive they acquire proficiency in the core competencies. An annual survey was sent to all residents at UCLA from 2007 to 2010. Survey questions asked trainees across various programs about the helpfulness of specific predefined learning activities in acquiring the competencies. Responses from 1378 PGY1-3 residents in 12 ACGME-accredited residency programs were analyzed. Patient care activities and observation of attendants and peers were listed as the two most helpful learning activities for acquiring all six core competencies. The findings reinforce the importance of learning from role models during patient care activities and the heterogeneity of learning activities needed for acquiring all the competencies. The fact that competencies are multidimensional and interconnected makes it highly unlikely that a single approach to teaching or assessment will be sufficient for their acquisition. Hence, multiple methods for teaching and learning are necessary for the acquisition of the competencies.

Takeaways

- Observation of peers and attendings in action, and patient care interactions were highly rated educational activities, reemphasizing the importance of proper role modeling in the clinical learning
- The fact that competencies are multi-dimensional and interconnected makes it highly unlikely that a single approach to teaching or assessment will be sufficient for their acquisition of the competencies
- Residents in our study obtained most of their knowledge about how to be a physician through their clinical activities and "by doing."

P.F. Wimmers (✉) · L.R. Guerrero · S. Baillie
David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
e-mail: pwimmers@mednet.ucla.edu; pwimmers@ucla.edu

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_6

6.1 Perspectives/Theoretical Framework

Six core competencies in medicine (patient care, medical knowledge, practice-based learning and improvement, interpersonal communication skills, professionalism, and systems-based practice) were implemented in 1999 and are now considered a fundamental part of professional training. The core competencies, as defined by the Accreditation Council for Graduate Medical Education (ACGME, 2011), offer a conceptual framework to address the knowledge and skills needed by students in training and doctors to perform competently (Nasca et al. 2012).

In order to meet the requirements for accreditation, residency programs are responsible to incorporate the core competencies into their training programs and identify learning activities to support trainees' acquisition of these. Yet, various authors have pointed to the difficulty in knowing how well residents have acquired a competency and how these can be effectively taught (Caverzagie et al. 2008; Cogbill et al. 2005; Lurie et al. 2009). Some have incorporated teaching the competencies into specific educational activities, like the teaching of practice-based learning and improvement into morbidity and mortality conferences practice (Fussell et al. 2009) or by providing seminars on professionalism and communication skills (Hochenberg et al. 2010). Residency program directors and GME offices have created specific curricular interventions or institution-wide didactic sessions. Although attention has been paid to the teaching, measurement and assessment of these competencies, less attention has been paid to learners' perceptions of how adequately they feel they are learning these competencies and which learning activities are most helpful.

Core competencies and related learning objectives are considered as educational outcomes and medical residents are required to demonstrate sufficient proficiency in all of these competencies independent of their residency. This means that professional training is primarily driven by output measures (objectives and competencies) rather than input measures (instruction and learning activities). An outcome is "what" you expect your residents to achieve. There is assumed that assessment based upon the core competencies provides evidence of the program's effectiveness in preparing residents for practice. The question of interest for educators is "how" residents perceive they acquire proficiency in the core competencies; the means to that end. The purpose of this study was to collect information on residents' perceptions of what learning activities contribute in acquiring the competencies. The specific question "Which specific learning activities have been most helpful in acquiring this competency?" has been asked via an annual resident survey since 2007.

6.2 Methods

This study analyzed 1378 responses from all years of the survey (2007–2010), across 3 years of training, PGY1 (*n* = 468), PGY2 (*n* = 450), and PGY3 (*n* = 460) from residents in 12 programs, including anesthesiology, emergency medicine, family medicine, head and neck surgery, internal medicine, medicine/pediatrics, obstetrics and gynecology, orthopedic surgery, pathology, pediatrics, psychiatry, radiology, and surgery. Residents were surveyed about which educational activities have been most helpful in their learning of the six core competencies. The educational activities listed included: (1) patient care interactions, (2) resident didactic teaching sessions, (3) journal club, (4) quality improvement sessions, (5) observation of peers and attendants, and (6) independent reading and study. These questions were scored on 5-point Likert scales (1 = not helpful to 5 = most helpful). Descriptive statistics (means and SDs) were used to analyze differences within and between programs. This study was approved by the UCLA Institutional Review Board, #10-000986.

6.3 Results

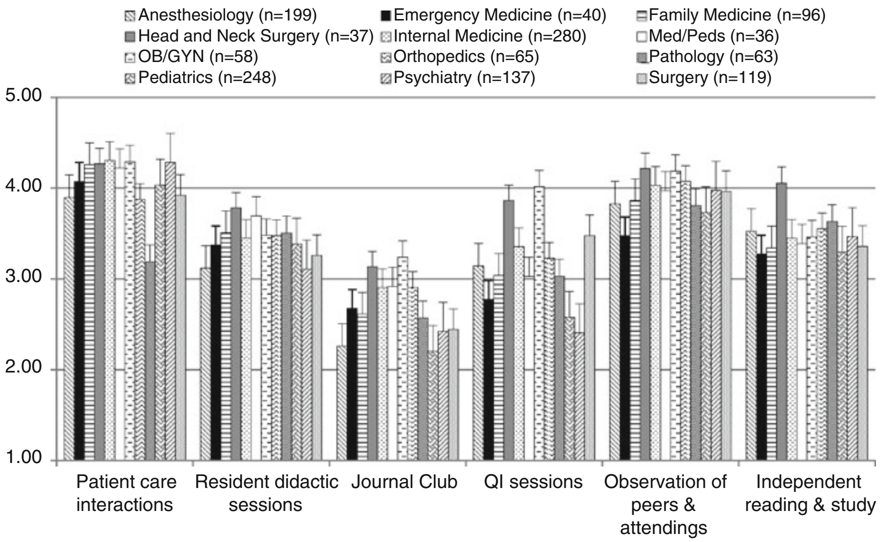
The overall response rates for this survey varied by survey year: 2007, 77 % (744/967); 2008, 81 % (755/933); 2009, 66 % (636/962); 2010, 82 % (812/989). Although, the degree of perceived helpfulness for each educational activity varied per competency, descriptive statistics show that all educational activities contribute to learning each individual competency.

Table 6.1 and Figs. 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6 portray the mean resident ratings of learning activities in terms of their helpfulness (1 = not helpful; 2 = slightly helpful; 3 = helpful; 4 = very helpful; 5 = most helpful) in acquiring the

Table 6.1 Average ratings of learning activities by ACGME competencies

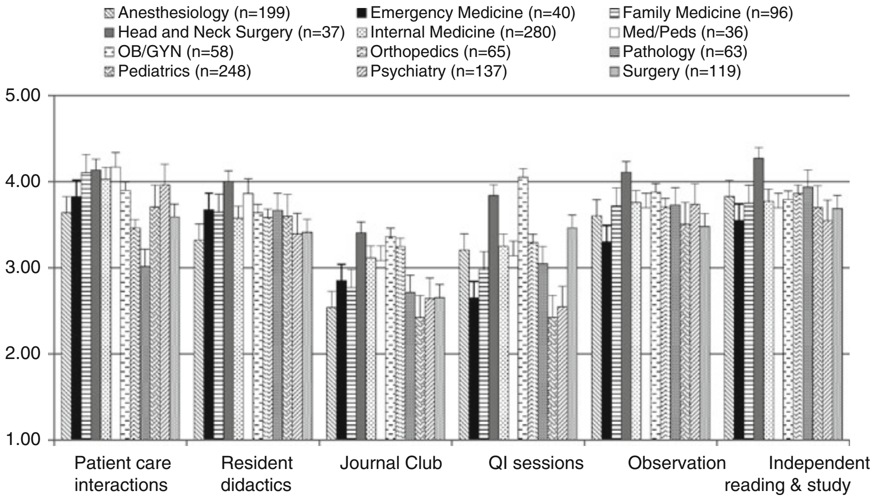
	Patient care interaction		Resident didactics		Journal club		QI		Observation of peers and attendings		Independent reading and study	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Patient care	4.08	0.888	3.36	0.865	2.58	1.02	3.08	1.03	3.91	0.896	3.44	0.902
Medical knowledge	3.80	0.930	3.54	0.894	2.79	1.04	3.05	1.06	3.66	0.932	3.75	0.930
Practice-based learning	3.42	0.101	3.13	0.968	2.71	1.07	3.14	1.08	3.36	0.991	2.93	1.09
Interpersonal communication	4.01	1.03	2.74	1.11	2.11	1.17	2.57	1.16	3.93	0.993	2.56	1.15
Professionalism	3.92	0.988	2.73	1.12	2.12	1.18	2.72	1.16	3.98	1.04	2.46	1.19
Systems-based practice	3.46	1.08	3.01	1.09	2.37	1.15	2.92	1.09	3.40	1.06	2.77	1.16

N = 1378



(Scale: 1=not helpful 2=slightly helpful 3=helpful 4=very helpful 5=most helpful)

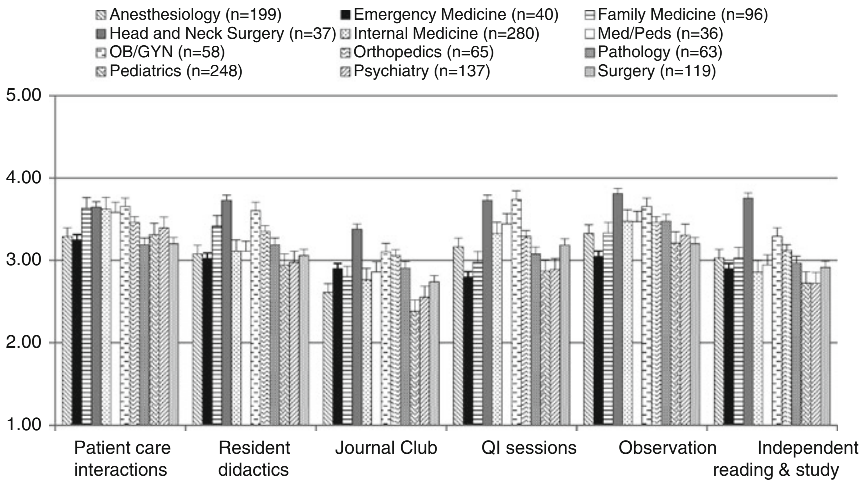
Fig. 6.1 Helpfulness of educational experiences for acquiring patient care competency



(Scale: 1=not helpful 2=slightly helpful 3=helpful 4=very helpful 5=most helpful)

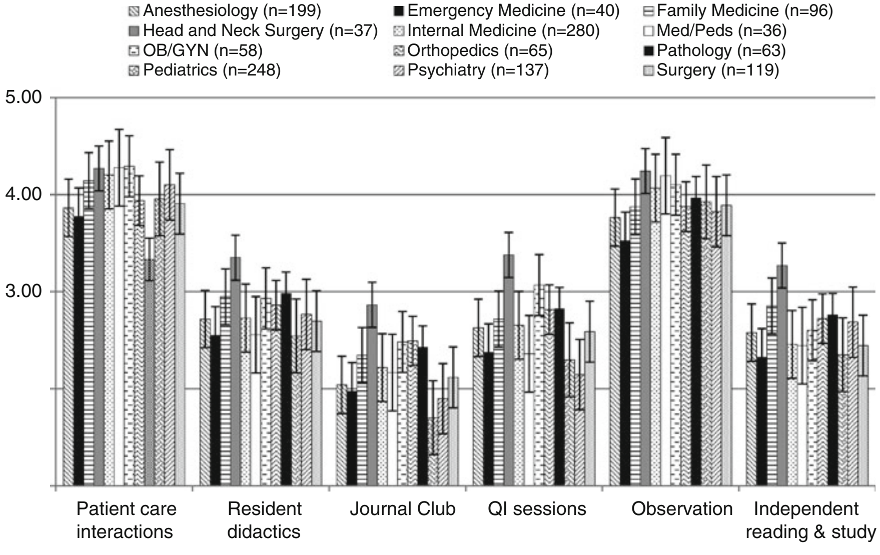
Fig. 6.2 Helpfulness of educational experiences for acquiring medical knowledge competency

six core ACGME competencies (bars represent standard deviations). The residents rated the patient care interactions as "most helpful" in acquiring five of the six competencies: patient care (M = 4.08; SD = 0.888), medical knowledge (M = 3.80;



Scale: 1=not helpful 2=slightly helpful 3=helpful 4=very helpful 5=most helpful

Fig. 6.3 Helpfulness of educational experiences for acquiring practice-based learning competency



Scale: 1=not helpful 2=slightly helpful 3=helpful 4=very helpful 5=most helpful

Fig. 6.4 Helpfulness of educational experiences for acquiring interpersonal and communication skills competency

SD = 0.930), practice-based learning and improvement (M = 3.42; SD = 0.101), interpersonal and communication skills (M = 4.01; SD = 1.03), and systems-based practice (M = 3.46; SD = 1.08). Observations of peers and attending were rated as

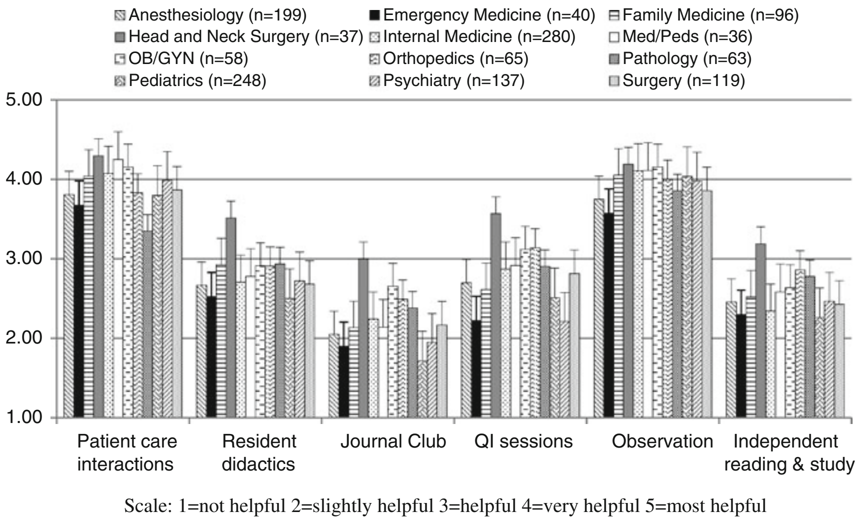


Fig. 6.5 Helpfulness of educational experiences for acquiring professionalism competency

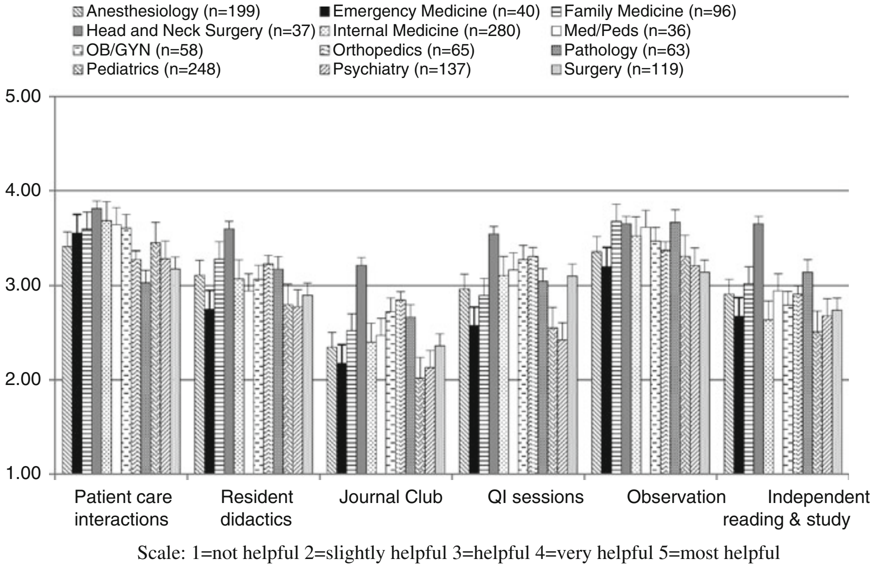


Fig. 6.6 Helpfulness of educational experiences for acquiring systems-based practice competency

"most helpful" in acquiring the professionalism competency (M = 3.98; SD = 1.04). For the acquisition of medical knowledge, independent reading received slightly higher ratings than patient care interactions among some of the specialties. Didactic teaching, a core teaching method of most residencies, was not rated highly. None of the activities received an average rating of "very helpful" in acquiring either the practice-based learning and improvement or the systems-based practice competencies. Last, observation of peers and attending in action, and patient interactions were highly rated learning activities, especially for professionalism and interpersonal communication.

Although the six core competencies have been defined as separate entities, our results show significant overlap among the educational activities that foster the competencies. This makes it difficult to assess competencies as independent constructs.

6.4 Conclusion

Specific educational activities foster multiple competencies. Competencies are not perceived to be learned through any single learning activity. The most helpful learning activity for the acquisition of medical knowledge was patient care interactions, although independent reading and resident didactic sessions were perceived as valuable too. For acquisition of the practice-based learning on the other hand, independent reading and study may be a more helpful way of learning this competency, rather than journal club. The self-reflection and analytical skills needed for this competency may not be gained in the clinical setting, suggesting that time for self-study and reflection might need to be formally built into trainee's schedules. Journal club's lower helpfulness ratings compared to other activities is surprising given one of the goals is gaining new medical knowledge. These findings may reflect another goal of these sessions which emphasize critical appraisal rather than knowledge acquisition. Similarly, interpersonal communication and professionalism are competencies perceived to be learned through interacting with patients and by observation of role models, not by independent reading or through journal club. Having residents work with a variety of patients in different settings seems to be the most helpful in acquiring the systems-based practice competency. This was somewhat surprising given the goals of QI sessions are often to look at systems problems and errors.

The fact that competencies are multidimensional and interconnected makes it highly unlikely that a single approach to teaching or assessment will be sufficient for their acquisition. Hence, multiple methods for teaching and learning are necessary for the acquisition of the competencies. Finally, we found that observation of peers and attendants in action, and patient care interactions were highly rated educational activities, this is especially interesting for professionalism and interpersonal communication (Cruess and Cruess 1997). This supports the literature which advocates the importance of proper role modeling in the clinical learning process (Cruess et al. 2008; Van der Vleuten and Swanson 1990). Similar to other studies, the residents in our study obtained most of their knowledge about how to be a physician through their clinical activities and "by doing."

Outcome-based (or competency-based) education can serve as a roof on a well-designed educational system but can never be a substitute of an ill-defined educational system that does not adequately address the means to achieve the outcomes. A pure focus on individual outcomes will deny the art in medicine and such an approach will only touch the surface of performance and not the depth and breathe of being a physician. Clinical competence takes place at the intersection of a lot of different learned abilities and skills. This ability of implementing and applying multiple core competencies is what medicine is about. Instruction and assessment are very closely related although they seem different.

Issues/Questions for Reflection

- Which learning activities have been most helpful in training students to be competent in a specific trade or profession?
- Clinical educators have to reflect on how they can more closely align teaching with assessment
- Clinical competence takes place at the intersection of a lot of different learned abilities and skills. Educators have to consider the implications for how it has to be measured or assessed to ensure their proper acquisition

References

ACGME. (2011). Outcome project: The common program requirements (Publication). Retrieved July 20, 2011 from http://www.acgme.org/outcome/comp/compfull.asp

Caverzagie, K., Shea, J., & Kogan, J. (2008). Resident identification of learning objectives after performing self-assessment based upon the ACGME core competencies. *Journal of General Internal Medicine, 23*(7), 1024–1027.

Cogbill, K. K., O'Sullivan, P. S., & Clardy, J. (2005). Residents' perception of effectiveness of twelve evaluation methods for measuring competency. *Academic Psychiatry*, *29*, 76–81.

Cruess, S. R., & Cruess, R. L. (1997). Professionalism must be taught. *BMJ, 315*(7123), 1674–1677.

Cruess, S., Cruess, R., & Steinart, Y. (2008). Role modelling making the most of a powerful teaching strategy. *BMJ, 336*, 718–721.

Fussell, J., Farrar, H. C., Blaszak, R. T., & Sisterhen, L. L. (2009). Incorporating the ACGME educational competencies into morbidity and mortality review conferences. *Teaching and Learning in Medicine: An International Journal, 21*(3), 233–239.

Hochenberg, M., Kalet, A., Zabar, S., Kachur, E., Gillespie, C., & Berman, R. S. (2010). Can professionalism be taught? Encouraging evidence. *American Journal of Surgery*, *199*(86–93).

Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. *Academic Medicine, 84*(3), 301–309 310.1097/ACM.1090b1013e3181971f3181908.

Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system—Rationale and benefits. *New England Journal of Medicine, 366*(11), 1051–1056.

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58–76.

Chapter 7
Performance Assessment in Legal Education

Erika Abner and Shelley Kierstead

Abstract Law schools in Canada and the United States are in a period of transition. New accreditation standards require law schools to ensure that students have acquired a range of competencies upon completion of their law degree. The articulation of competencies, like Carnegie's emphasis on the three apprenticeships to which law schools should strive, invites greater scrutiny into curriculum design, teaching methods, and student assessment. Law schools are shifting from traditional case law teaching and 100 % final examinations in doctrinal courses to heightened focus on experiential learning and performance assessment—approaches that have already been adopted in clinical legal education, legal writing, and skills-based courses. In this chapter, we conducted a literature review on performance assessment that included the legal and medical education literature, websites of legal education organizations, and of regulatory and licensing bodies. In Canada, under each provincial body, many bar admission courses have incorporated performance assessments into their programs. However, this has neither been uniform nor have the initiatives been rigorously evaluated. While more law schools are adopting performance assessment methods, disciplined theory-based inquiry proceeds slowly. Legal academics rarely refer to the leading authors on measurement and evaluation or to the extensive literature from other professions, notably medical education. With modest exceptions, these authors do not address established conventions: (a) alignment of instructional objectives to assessment methods, (b) processes for development of assessment instruments, (c) analysis of validity and reliability, and d) reporting of results. As a result, it is difficult to ascertain whether a particular course or program has achieved any or all of the five elements for faculty and student learning: (1) Self-reflection on learning own abilities; (2) Self-assessing performance and using feedback to improve it over time; (3) Learners developing metacognitive performance; (4) Learners developing

E. Abner (✉)
Faculty of Medicine, University of Toronto, Toronto, ON, Canada
e-mail: erika.abner@utoronto.ca

S. Kierstead
Osgoode Hall Law School, Toronto, ON, Canada
e-mail: skierstead@osgoode.yorku.ca

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_7

professional expertise; and (5) Learners developing identity as a self-sustained and unique learner, contributor, and professional. To effect real change in performance assessment that incorporates these elements, we argue that law schools need a robust core of trained specialists and comprehensive faculty development in performance assessment, program assessment, measurement, and evaluation.

Takeaways

- Legal educators are at an early stage in their development of programmatic assessment; relatively few educators are trained in educational research.
- Law teachers are engaging in creative performance assessment; however, there appear to be few opportunities for a standardized approach to knowledge production. That is, the academy would benefit from rigorous public discussion of the development, use, and validation of assessment instruments. Law schools should report on assessment as a feature of any curriculum renewal or change.
- Individual teachers would benefit from more faculty development in performance assessment. In particular, adjunct teachers (often practitioners) who work in relative isolation value support within their community of practice.

7.1 Introduction

This chapter describes a case study of performance assessment at Osgoode Hall Law School, which is one of Canada's largest law schools. Legal education is undergoing significant changes in philosophy and practice, following the publication of the influential *Educating Lawyers* (Sullivan et al. 2007) and *Best Practices in Legal Education* (Stuckey 2007) and against the background of the focus on outcomes assessment in accreditation standards in both countries. These changes include more focus on experiential learning, developing "practice-ready" graduates, and demonstration of performance through assessment. Performance in legal education includes a wide range of knowledge, skills, and abilities: legal analysis and problem solving, interviewing and counseling, negotiation, writing and drafting, and advocacy. Law schools, including Osgoode Hall, have developed a wide range of courses and experiences together with assessments shaped to the performance.

Understanding performance assessments in legal education requires an initial understanding of the context, place of learning (classroom or clinic), focus (knowledge, skills, values), and teacher (doctrinal, clinical, writing, other). As described below, law school courses can be loosely divided into six major educational settings, of which five offer skills education and may include performance assessments. The first, traditional category is the most familiar: the doctrinal classroom—generally large group format—employing Socratic dialog to develop a specific legal area, such as torts or criminal law. These classes are distinguished by the use of single point end of term examinations that assess issue spotting, organization, and analysis. Small group seminars may also use discussions and presentations as a teaching method and rely on course papers for evaluation.

However, law schools now include a broader array of contexts, places, and topics of learning. These other educational settings have been described as:

1. Those that allow students to gain real case experience, such as in-house clinics or externships;
2. Skills-focused simulation courses, such as Interviewing or Negotiations;
3. Practice context simulation sources, such as Criminal Practice or Real Estate transactions;
4. Doctrinal courses with skills exercises as part of the pedagogy, such as Family Law, where students conduct lawyering exercises based on a simulated divorce problem; and
5. Doctrinal courses that critically analyze lawyering[1] (Katz 2008, p. 924)

These settings generally include some form of performance assessment, often paired with a reflective component. The momentum for increased levels of experiential learning continues, as institutions share experiences through conferences and online reports. Law schools in both the United States and Canada are expanding the depth, breadth and sophistication of the experiential learning element of law school curricula (Experiential Learning Symposium, 2014).

While shared experiences are important, scholarly attention to performance assessment has been fragmented. Performance assessments in legal education generally have not been analyzed from either a psychometric perspective (for example, psychometric properties, validity, cost and feasibility) or a programmatic perspective. This case study seeks to provide analysis and theory testing of a particular outcomes framework in legal education, that of Mentkowski et al. It asks the question:

Do performance assessments in legal education include, either implicitly or explicitly, reference to the following outcomes, learning processes, and transformative learning cycles:

1. Self-reflection on own learning abilities: a faculty-stimulated learning process of using self-assessment to engage learning in observing their performance of abilities.
2. Self-assessing performance and using feedback to transform it over time: a learning process of deepening one's learning by interpreting, analyzing, and judging one's performance, gradually transforming it over time with diverse feedback from trained assessors.
3. Learners develop metacognitive performance: a transformative learning cycle that assists learners to engage in restructuring their knowledge. Learners

[1] Erika Abner LLB LLM Ph.D. is the Faculty Lead, Ethics and Professionalism, Faculty of Medicine, University of Toronto [erika.abner@utoronto.ca]. Shelley Kierstead LLB LLM DJur is an Assistant Professor and Director, Legal Process, Osgoode Hall Law School, York University [skierstead@osgoode.yorku.ca]; both of the Ontario Bar.

recognize patterns in knowledge-rich environments, thinking while performing and thinking about disciplinary frameworks.

4. Learners develop professional expertise, as they engage in reflective learning, envision improved role performance, and monitor role performance in relation to criteria and standards across multiple, varied situations to reach beyond what they thought they could do.
5. Learners develop identity as a self-sustained and unique learner, contributor, and professional: a learning process of engaging in learning independently, somewhat before they engage in and accomplish a developmental restructuring of their thinking and reasoning (Mentkowski et al. 2013, pp. 29–30).

We note an important element within this framework, which is the capacity to render judgments with integrity under conditions of both technical and ethical uncertainty. This outcome is derived from Schön's portrayal of the reflective practitioner (1983, 1987, 1995) as well as Gardner and Shulman's overarching vision of professional work: "the essential challenges of professional work center on the need to make complex judgments and decisions leading to skilled actions under conditions of uncertainty. This means that professional practice is frequently pursued at or beyond the margins of previously learned performances" (2005, p. 15).[2] In law, these "conditions of uncertainty" have been described as the "five indeterminacies of lawyering; learning lawyering is about managing these indeterminacies both individually and collectively." These five indeterminacies include not only law and facts, but also problem solving, lawyering skills and the lawyer–client relationship (Dinerstein and Milstein 2014, p. 328).

We provide two cautionary notes to any attempt to fully understand curriculum and performance in legal education. First, research into assessment in legal education can be challenging. It is not uncommon for legal academics to cite only to other legal academics for concepts originally developed and disseminated by education academics. With relatively few legal academics trained in research (either qualitative or quantitative) and even fewer psychometricians, research into performance assessment is very limited and what is available is rudimentary. Unlike medical education, with numerous journals devoted to teaching, learning and assessment, reports and studies on legal education can be difficult to locate. Finally, relatively few assessment concepts from other professions have been adopted or even acknowledged in legal education. So, for example, concepts such as programs of assessment (Dijkstra and van der Vleuten 2010) or entrustable professional activities (ten Cate 2013) do not inform current curriculum and assessment renewal efforts. There are limited references in the literature to assessment instruments that might be transferable into certain contexts, such as the mini-CEX and other forms of workplace assessment (Norcini and Birch 2007). Finally, only a few studies attempt disciplined analysis of the learning outcomes of curricular experiments or innovations.

[2] These outcomes will be described collectively as the "Mentkowski outcomes".

Second, many schools are implementing curricular and assessment changes and reporting on these changes through conferences, websites, and non-peer reviewed databases such as the Social Science Research Network [ssrn.com]. These changes are occurring with some rapidity, so that statements made today may well be out of date in the near future.

This chapter is divided into four parts. The literature review in Sect. 7.2 includes a review of the nature of lawyers' performance and surveys the available literature on performance assessment in legal education across the educational settings as described above. Section 7.3 describes the methods used in the literature review and the case study undertaken at Osgoode Hall Law School of York University. Section 7.4 reports on the case study, which examines the nature and extent of performance assessment at Osgoode following its First Year and Upper Year Curriculum Review Reports from three different perspectives:

1. an analysis of syllabi of the Osgoode courses that fit within Katz' categorization of educational settings,
2. an analysis of interviews of four law professors who offer praxicum courses at Osgoode, and
3. an in-depth review of a single praxicum course.

As we examine performance assessment in detail through these different perspectives, we discover the interplay among theory, practice, and reflection in the students' development as professionals. The discussion and conclusions are contained in Sect. 7.7.

7.2 Literature Review

7.2.1 Performance in Law

Extensive frameworks of lawyering knowledge, skills, tasks, and attributes have been developed over the past 35 years in different jurisdictions (Cort and Sammons 1980; Fitzgerald 1995). More recently, frameworks have been developed for national bar admission examinations (Case 2013a, b; Federation of Law Societies of Canada 2012, 2015), law school admissions (Shultz and Zedeck 2011), and for law firm performance management systems (Bock 2006; Chitwood and Gottlieb 2000; Sloan 2002). These frameworks continue to be expanded as academics and practitioners identify additional skills desirable or necessary in practice (Daicoff 2012).

While the four prominent frameworks described below were created by different methods (surveys, interviews and focus groups, expert panels, job journals) and with different objectives, they share some common characteristics. First, these frameworks tend to be analytical rather than synthetic (ten Cate 2013; ten Cate and Scheele 2007) and linear (with some exceptions noted below)—raising a concern that competencies may be assessed in isolation. Second, the frameworks tend to be underinclusive: (a) not acknowledging current competencies, such as legal research

(Alford 2009) and (b) not including new competencies, such as use of technology. Finally, the frameworks may fail to take into account knowledge use in context—how legal analysis and problem solving is actually accessed in practice (Eraut 1994; Roper 1999).

Despite these critiques, frameworks that describe lawyering competencies play a central role in performance assessments. First, descriptions of competencies form the basis for the test specifications of knowledge, skills, and attitudes together with tasks (Johnson et al. 2009, pp. 35–37). Second, these frameworks are essential for professions shifting to outcomes-based education, which "starts with a specification of the competencies expected of a physician, and these requirements drive the content and structure of the curriculum, the selection and deployment of teaching and learning methods, the site of training, and the nature of the teachers" (Norcini et al. 2008, p. 2). As developed in Sect. 7.4, certain frameworks appear to be more explicitly employed at the level of bar admission; it is not clear the extent to which law schools utilize any particular framework.

These competency frameworks can be divided into four separate categories.

A. *Knowledge Frameworks* have been developed by the Law Society of Upper Canada for use in creating its Bar Admission examinations. The *Competency Standards for Entry-Level Barristers* and the *Competency Standards for Entry-Level Solicitors* (Professional Development Competence and Admissions Committee 2004) set out separate competency categories and subcategories for solicitors and for barristers. Since the profession is fused, an entry-level lawyer must be competent across the entire range. While the categories are quite broad, the substance is detailed. For example, the first category, "Ethical and Professional Responsibilities" includes 24 separate items. The Real Estate category alone, under the general heading of Knowledge of the Law, states that the new lawyer must "demonstrate knowledge of substantive real estate law including the following primary statutes and related regulations and case law," and sets out seventeen primary statutes and 21 secondary statutes (and that is included in the first two of 30 categories of knowledge in the real estate section alone.).
B. *Skills and Knowledge Frameworks* have recently been developed by the National Conference of Bar Examiners, through an extensive workplace tasks analysis, as well as the Federation of Law Societies of Canada. Both frameworks are being analyzed for use in bar admission examinations.
C. *Attributes, Values, and Skills Frameworks* have been developed by the American Bar Association (American Bar Association Section of Legal Education and Admissions to the Bar 1992) and more recently, through independent research conducted by an organizational psychologist and a law professor (Shultz and Zedeck 2011). While developed through different methods nearly twenty years apart, these two frameworks are remarkably similar at the conceptual level.
D. *Integrated Frameworks* have, not surprisingly, taken hold in law firms, where performance assessment is framed as a human resources function and grounded in workplace activity. Unlike the competency frameworks described above, law

firm frameworks focus on described levels of development, often incorporating concepts of increased independence (ten Cate 2013). Competency frameworks are created within the law firms, and therefore described as consistent with individual firm culture (Manch 2010). These frameworks are often not linear, and may employ various visual elements (similar to the CanMEDS "flower") to emphasize central concepts and embedded values, for example, that described at the law firm Reed Smith (http://www.reedsmith.com/career_development/).

7.2.2 General Issues of Performance Assessment in Legal Education

This section provides a brief overview of current issues in performance assessment: scholarly concerns, the focus on curricular review without corresponding focus on assessment, the shortage of educational research, and limited detail provided in the literature on assessment practices.

Educating Lawyers addresses the issue of assessments in legal education. The work generally describes three "apprenticeships" that legal education should provide: (1) cognitive reasoning; (2) skills and practice; and (3) professional identity and values. In addressing assessment, the authors note that there is significant focus on assessment of the first apprenticeship, and that "significant work is needed to establish widely used, highly valid procedures for assessing the skills and qualities of the practice and ethical-social apprenticeships" (Sullivan et al. 2007, p. 175). *Best Practices* includes prescriptive detail in the form of an extensive list of recommended assessment practices that distinguishes formative from summative assessment, employs criterion-referenced marking, and uses multiple methods (Stuckey 2007, Chap. 7).

Scholars continue to express concerns over the role of assessment in legal education. Wegner (2009, 2013) supports the argument that assessment drives learning and advocates for "fresh assessment strategies to challenge assumptions" (2013, p. 63). Hertz et al. (2013) identify continuing barriers to curriculum renewal, including the cost of experiential education, the lack of expertise in law school faculties in outcomes assessment, and the potential for assessing only what is quantifiable rather than what is desirable. Lynch (2011) addresses professorial concerns about outcomes evaluation, including such issues as teaching to the test, impinging on academic freedom, and student resistance. Progress to understanding assessment practices is hampered by legal academics skeptical of the value of educational research.

Law schools have increased all aspects of skills instruction, including clinical, simulation, and externships; the greatest growth in professional skills courses have been in Transactional Drafting and upper level Legal Writing. Over 85 % regularly offer in-house live-client clinical opportunities, with 30 % offering off-site, live-client clinical opportunities. Data from the 2012 Report of the Law School

Survey of Student Engagement ("Law School Survey of Student Engagement: Lesson from Law Students on Legal Education," 2012, pp. 14–15) confirm relatively high rates of participation in experiential courses such as clinics, externships and skills courses. The Survey report concludes that experiential learning contributes to student perceptions "that their law school classes emphasize higher order learning activities, including analysis and synthesis of ideas and information, making judgements about the value of information, and applying theories and concepts to practical problems or in new situations."

However, this trend toward reporting of curriculum redesign in the form of an expanded array of experiential courses is not accompanied by **reporting** of integrated assessment methods. For example, the *2010 Survey of Law School Experiential Learning Opportunities and Benefits Report* (National Association for Legal Career Professionals and The NALP Foundation for Law Career Research and Education 2011) surveyed students regarding courses, clinics, and pro bono work, but contains no information on assessment. A similar survey conducted by the American Bar Association (Carpenter 2012) concluded that curricular review "has produced experimentation and change at all levels of the curriculum, result in new programs and course, new and enhanced experiential learning, and greater emphasis on various kinds of writing across the curriculum." Of the nine law schools described in *Reforming Legal Education: Law Schools at the Crossroads* (Moss and Curtis 2012b) only three provide detail on performance assessments as key to curriculum reform. Further, assessment is not included in the chapter on essential elements for the reform of legal education (Moss and Curtis 2012a). As noted by Engler (2001, p. 144) regarding our understanding of assessment practices in law schools: "assessment is complicated further by difficulties in speculating about the impact on particular law schools of general trends or, alternatively, collecting information one law school at a time and determining a cumulative impact."

Despite the general and specific concerns raised about assessment, law schools do have a history of different forms and creative approaches. Works such as *Outcomes Assessment for Law Schools* (Munro 2000) and the *Moving Students from Hearing and Forgetting to Doing and Understanding* manual (Ramy 2013) contain basic overviews of assessment design and examples of assessments of different skills. The Institute for Law Teaching and Learning website contains a variety of information on assessment. *Techniques for Teaching Law*, a pivotal text, advises that "[E]ffective evaluation schemes of adult learners have three characteristics: multiple, varied, and fair" (Hess and Friedland 1999, p. 289) and describe assessments such as a video examination that evaluates oral lawyering skills (308–310), extemporaneous oral examinations (313), and attendance/reporting on an administrative hearing (128). Post-Carnegie publications, such as *What the Best Law Teachers Do*, describe the importance of designing assessments that are seen as challenging, as well as fair and transparent: "[T]hey intentionally design their assessments to be congruent with course learning goals" (Schwartz et al. 2013, p. 280). Professors use rubrics, portfolios, multiple forms of feedback, and various forms of writing assignments and practice examinations to ensure continuing understanding and development. The 2010 Clinician's Conference on Outcomes

Assessment (AALS Conference on Clinical Legal Education 2010) describes innovative approaches in clinical legal education. For these examples, and generally, performance assessment is reported in broad detail with limited discussion of development and analysis. A search of all issues of the Clinical Law Review found very few articles that included actual assessment instruments. Some articles, described in Sect. 7.4, include assessment instruments such as rubrics, instructions for reflective journals, or externship reports.

Earlier studies have found a limited research base that informs legal education (Neumann and Krieger 2003; Ogloff et al. 2000). Despite the energy and creativity described above, research into assessment continues to be underdeveloped with few clear lines of inquiry. Sargent and Curcio (2012, p. 395) examined the effect of providing formative feedback in a large doctrinal course [Evidence] in the form of ungraded quizzes and a graded midterm, finding that "70 % of the intervention group benefitted substantially (nearly a letter grade) from the formative assessment materials." Lopez et al. (2009, p. 66) assessed the communication skills of medical students and law students engaged in client interviews using standardized patients. Using a pre and posttest design, following an educational intervention, the researchers found that the "experience provided opportunities to learn new information and skills" even though there were no statistically significant changes in communication skills. Barton et al. also examined communication skills in an OSCE-like exercise with Standardized Clients (Barton et al. 2006). Two small studies examine the effects of different grading regimes in live-client clinics (Brustin and Chavkin 1996; Murray and Nelson 2009). Finally, Stefan Krieger has engaged in a series of studies designed to examine the effect of clinic participation on students' legal reasoning abilities (Krieger 2008, 2011).

7.2.3 Specific Examples of Performance Assessment

This section provides a brief overview of different types of performance assessments across the different legal education settings.

Doctrinal. Courses that focus on specific areas of the law have traditionally used a single point end of term examination as the sole method of assessment (Kissam 1989; Sheppard 1997). Recent articles describe different approaches to assessment in doctrinal courses, including assessments that incorporate advanced problem-solving skills as well as professionalism and ethics. For example, Curcio (2009) advises including activities such as fact investigations, review of video simulations, drafting, and oral performance exercises. Sergienko (2001) describes self-assessment in addition to final examinations and recommends the use of multiple choice questions to test skill development. Corrado (2013) describes the development of a midterm examination in his Administrative Law and Contracts classes, intended to provide formative feedback through use of an extensive rubric and posted exemplars. Students are encouraged to appeal their midterm mark; the activity of preparing a one-page argument is described as further use of analytical skills.

Two studies provide more detail on experiential learning in doctrinal courses. Friedland (2013) has created an Evidence (traditionally considered doctrinal) course that includes applied Trial Advocacy (traditionally considered a skills course); the students experience simulations, field activities, writing exercises, and mock trial participation. Grose (2012–2013) provides a careful overview of the process of designing, incorporating, and evaluating outcomes-based assessment in a Trusts and Estates course.

Skills. Skills courses such as interviewing and counseling, negotiations, legal drafting and trial advocacy enjoy a rich tradition of creative approaches to performance assessment. A recent book (Ebner et al. 2012) on negotiations courses includes entire chapters on different forms of performance assessment. These range from the traditional "blue book" examinations through to the use of a "reputation index" and include quizzes, role plays, interviews, video recordings, and negotiation competitions. As well, a review of the syllabi of negotiation courses in different law schools found that 15 out of 19 courses used journals or some other forms of written reflection as part of the grading system. The use of the Reputation Index is particularly interesting; Welsh describes how she requires students, for 5 % of their course mark, to complete a Reputation Index toward the end of term. This tool "permits students to nominate other students in the class who they perceive have achieved the most positive or negative reputations as legal negotiators" (Welsh 2012, p. 175).

Grosberg (2006) provides an overview of assessment of interviewing and counseling skills in the clinical context, noting that the traditional approach has been for clinicians to grade these skills by direct observation. Grosberg describes assessment tools that he has used over time and that could supplement direct observation: videotaped performance test (a written exam based on analysis of the video scenario (s), multiple choice questions (possibly based on a transcript of a videotaped interview that the students had viewed), self-reflection (written or oral), standardized clients, and computerized assessment tools that incorporate videotaped performance tests and multiple choice tests. He concludes that clinicians should be open to using different measurement tools than those "used to assess students' grasp of doctrinal law."

Binder et al. (2006) argue for the inclusion of a deposition practice course, since depositions form a critical role in civil litigation and employ complex skills. Their article describes the authors' experience with teaching in-class and live-client depositions courses; the former rely on simulations while the latter rely on simulations with live-client depositions occurring later in the term. Students undertake a number of different simulations designed to teach different aspects of depositions, with considerable feedback for each. Some exercises are videotaped and the students are expected to assess their own performance.

Legal Research and Writing. The focus on skills building in Legal Research and Writing (LRW) courses has meant that there is a history of varied performance assessment within the LRW field that has not typically existed in many doctrinal courses. The use of formative assessment by way of student conferences to review draft memos and briefs is frequently employed (Wellford-Slocum 2004), as are citation quizzes, guided self-graded drafts (Beazley 1997) and peer feedback

(VanZandt 2010). LRW teachers regularly use rubrics (Sparrow 2004; Clark and DeSanctis 2013) for assignments such as memos, opinion letters and court briefs. Student portfolios showing the progression of writing skills, and sometimes reflecting on the process undertaken to achieve that progression, are another form of assessment familiar to legal writing teachers (VanZandt 2010). In short, along with clinical programs, LRW teachers are innovators of different approaches to performance assessment. Documentation of these innovations is available in publications such as the Legal Writing Institute's newsletter, "The Second Draft" (2010, Volume 24, no. 3).

Clinical Legal Education. Clinical legal education includes simulation courses, live-client clinics, and externships (Stuckey 2006). Clinical legal education may include practice observation and/or client representation. Objectives of clinical legal education may vary widely, depending on the institution, and may focus on: legal analysis, training for practice, and development of a professional identity. Some clinical scholars describe the objectives as improving students' professional skills and imbuing students with the desire to devote their professional lives to legal and social reform (Binder and Bergman 2003).

Externships. Externships have been described as "a type of clinical experience in which a student works for academic credit in a legal setting outside the law school under the supervision of an attorney and also attends a related seminar class at the law school." (Terry 2009, p. 243). Terry describes the four basic components of the externship structure as: field placements (a wide variety of types of placements), journals, supervision, and seminar classes that discuss and reflect on experiences in the field. Reflective journals have been identified as central to externship pedagogies (Katz 1997; Terry 2009).

Law schools are expanding their externship programs (Backman 2013; Basu and Ogilvy 2012) as placements are less expensive than clinics but still provide some real world experience (Backman 2006). Externships have been studied extensively by clinicians; for example, the Clinical Law Review Volumes 14(1) 15(2), 19(1) are devoted entirely to externship pedagogy.

Live-Client Clinics. Live client clinics provide opportunities for students to interact directly with clients. Debates about the purpose of clinical legal education abound; different authors promote different objectives. For example, Binder and Bergman (2003) identify two overall goals: to improve students' professional skills and to develop students' intent to devote their lives to legal and social reform. Grose (2012, p. 493) describes clinical legal education as an opportunity to teach students to approach lawyering as a theory-driven practice: "providing learning for transfer; exposing students to issues of social justice; and offering opportunities to students to practice lawyering skills." Martin and Garth argue that clinical legal education offers an antidote to the "misery" embedded in the transition from school to practice, providing opportunities to acquire lawyering skills in law school (1994).

Menkel-Meadow (1986, pp. 289–290) examines theory-building in clinical legal education, which includes analysis and development of performance together with analysis and critique of the "relationship of doctrine and substantive law and process to the practice of lawyering skills…to understand the political, economic and social foundations and structures of the rule systems that our students will be working with when they are lawyers. To think about what the rule is and from where it is derived is an essential "skill" of being a lawyer."

Clinicians have written extensively on developing performance assessments that provide feedback on skills and on commitment to social justice (Cavazos 2010; Ziegler 1992). Two methods are prominent: reflective practice (Spencer 2012) and extensive critique "on several simultaneous levels—objectives, performance, analysis, social role, effect on others and learning…" (Menkel-Meadow 1986, p. 288). Leering (2014) specifically examines reflective practice through an action research project conducted at Osgoode Hall Law School, developing a sophisticated multidimensional concept of reflection.

Integrated Curricula. Law schools are beginning to report on their experiences with integrated curriculum. The University of New Hampshire School of Law has developed the Daniel Webster's Scholars Honors Program, an intensive two year program of required courses in second and third year. These courses include "practice courses that would be small, emphasize the MacCrate skills and values, and be taught in the context of real life." Students experience a combination of formative, reflective and summative assessment (Gerkman and Harman 2015).

In contrast, Touro Law School has developed a three-year program for all students, called Portals to Practice. Portals to Practice is a carefully designed and sequenced as "multidimensional apprentice-based education that integrates the core competencies"; students learn through observation and shadowing, simulations, hybrid courses, and externships. Third year is comprised of an immersive clinical experience.

Other schools have developed intensive experiential components in the first year curriculum, generally with live-client representation (Capulong et al. 2015).

7.3 Methods

Literature Review. The literature review employed a variety of different methods in addition to a standard review of legal education databases—LegalTrac, Index to Legal Periodicals, Social Science Research Network, and Canadian Legal Periodicals—using the search terms "assessment," "outcomes assessment," "performance assessment," and "evaluation." The table of contents of all issues of the Clinical Law Review, the Journal of the Legal Writing Institute, the Canadian Legal Education Annual Review, and the ALWD Journal of Legal Rhetoric and Writing were reviewed. Key articles were noted up through Google Scholar.

Articles were selected based on:

1. Direct research, including action research, on some form of learner assessment (including assessment methods in classrooms and clinics).

2. Essays on assessment and evaluation in legal education.
3. Reviews or discussions of current issues in legal education.

Because much of the work in legal education is now shared through websites and blogs, the following were reviewed for any research, reports, or other current information:

1. American Bar Association Section on Legal Education and Admissions to the Bar: https://www.americanbar.org/groups/legal_education.html
2. American Association of Law Schools: http://www.aals.org/
3. National Conference of Bar Examiners: http://www.ncbex.org/
4. Federation of Law Societies of Canada: http://www.flsc.ca/
5. Law Society of Upper Canada: http://www.lsuc.on.ca
6. Law Society sites for all other provinces and territories, which can be found at: http://www.flsc.ca/en/canadas-law-societies/
7. The Alliance for Experiential Learning in Law (and a number of individual schools listed as members): http://www.northeastern.edu/law/experience/leadership/alliance.html
8. Best Practices in Legal Education: http://bestpracticeslegaled.albanylawblogs.org/
9. Institute for Law Teaching and Learning: http://www.lawteaching.org/

Case Study: Osgoode Hall Law School. Osgoode Hall Law School is situated within York University's campus in Toronto, Ontario. It has implemented two major curriculum reform initiatives within the past decade, along with several incremental changes. These reforms introduced a number of initiatives that open possibilities for more varied performance assessment. Osgoode has a standing tradition of offering various clinical programs which provide opportunities for robust assessment formats.

The sources for the case study consist of:

- Osgoode Hall Law School's First year Curriculum Reform Report (2007)
- Osgoode Hall Law School's Upper Year Curriculum Review Report (2010)
- Syllabi for Osgoode Hall Law School Clinical and Intensive Programs
- Syllabi for other courses that fall within Katz's categorization scheme
- Interviews with four instructors whose course offerings meet Osgoode's "praxicum" requirement
- An in-depth description of one of Osgoode's "praxicum" course offerings.

Syllabi were analyzed using NVivo Software, with the performance assessment modes set out in each syllabus being coded in accordance with Mentkowski's outcomes.

Instructor interviews were conducted for courses involving a range of substantive and lawyering skills coverage: a criminal law intensive program, a collaborative law seminar, a mediation seminar, and a corporate finance workshop. Instructors were asked to describe their teaching method(s) for the course, the nature of required student performance, and their approach to assessing performance. The interviews were analyzed using NVivo Software, with each element of the interview being coded to capture the range of teaching techniques and assessment methods. The responses were then analyzed through the lens of Mentkowski's outcomes. While the courses were chosen to reflect a range from "traditional" content focus (criminal and corporate finance) to established curricular skills offering (mediation intensive) to fairly new lawyering methods content (collaborative lawyering), the limitations of this study are readily acknowledged. The results cannot be generalized to other programs, but rather, are aimed at providing a starting point for further research.

7.4 Case Study: Osgoode Hall Law School

With a first year class of between 280 and 290 students, Osgoode is one of Canada's largest law schools. In its most recent Strategic Plan, Osgoode defined three important trajectories for the school: (1) toward experiential education and the exploration of law in action; (2) toward the intensification of research and pushing the bounds of legal knowledge; and (3) toward an engaged law school in the community, province, country and the world. The first trajectory, toward experiential education and the exploration of law in action, clearly calls for a robust approach to performance assessment. A brief explanation of Osgoode's major curriculum reform initiatives over the past decade provides a backdrop to the subsequent analysis of its current curricular offerings.

7.4.1 First Year Curriculum Reform

In 2007, the first year curriculum at Osgoode Hall Law School was revamped.[3] The key reforms achieved within this initiative were the move to the semestering of a number of courses, the merger of the Legal Research & Writing and Civil Procedure courses into a course titled "Legal Process," the creation of a course combining concepts of public law and constitutional law titled "State and Citizen," and the creation of a course called "Ethical Lawyering in a Global Community." For the purposes of this discussion, the two most significant initiatives were the creation of the Ethical Lawyering course and a focus within the Legal Process course on contextualized problem solving.

7.4.1.1 Ethical Lawyering

The Ethical Lawyering course was designed to introduce students to the types of ethical decision-making required of competent lawyers as they practice within

[3] A copy of the Report is on file with the author, Shelley Kierstead.

increasingly globalized contexts. The course was approved after intense deliberation amongst faculty about the best way to teach ethics within a law school. Ultimately, the rationale behind situating the Ethical Lawyering course within the first year program was to introduce students at an early stage to ethical considerations which they could then transfer into other courses and situations. Each instructor at Osgoode Hall Law School is required, pursuant to a Faculty Council directive, to incorporate ethical considerations within his or her course. In this way, students will have a foundation from which to consider ethical questions raised throughout the curriculum. The Ethical Lawyering course is offered in an intensive format for one week at the beginning of students' first academic year and for another two weeks at the beginning of the winter term of the same academic year. Evaluation consists of both written papers and simulated skills.

In 2006, Osgoode had adopted a public interest service requirement for all of its students. Through their three years at law school, students must undertake 40 hours of legal work that can be broadly framed as serving the public interest. Prior to graduation, students must either complete a reflective paper relating to their public interest work or engage in a three-hour group reflective session. Within the reflective session, students are often asked to make linkages between the subject matter of their public interest work and their learning from the Ethical Lawyering course.

7.4.1.2 Legal Process

The Legal Process course combines traditional legal research and writing topics with basic civil procedure topics and an introduction to alternative dispute resolution methods. The course focuses heavily on skills development, beginning with an understanding of the importance of statutory and case law reasoning, moving to an understanding of specific methods for researching legal problems, and then focusing on the drafting of specific documents required within typical legal practice. Students also engage in an oral advocacy exercise during which they take on the role of legal counsel in a mock court procedure. Many of the assignments are situated within the context of a case study aimed at contextualizing the skills learning.

7.4.2 Upper Year Curriculum Reform

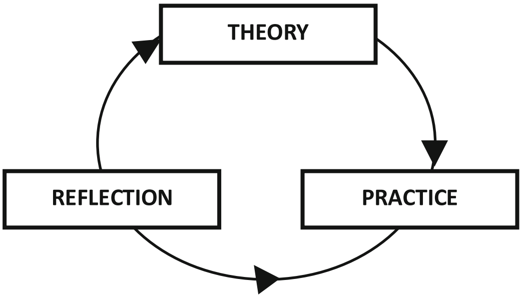
In 2011, Osgoode faculty adopted a set of upper year curriculum reforms.[4] The key changes to the curriculum resulting from this initiative were the creation of a praxicum requirement along with an advanced paper writing requirement. The purpose of the praxicum is to foster experiential learning within classes. It is intended to ensure that students who have not taken a clinical or intensive program

[4] The Upper Year Curriculum Reform Report is on file with the author Shelley Kierstead.

by their third year and will not be taking any of these programs during their third year will have at least one substantial opportunity to make specific connections between theory and practice.

A praxicum, as envisioned in the reformed upper year curriculum, is a seminar, course or program of study that actively integrates legal theory with practice. Further, by doing so, a praxicum allows students simultaneously to engage with and reflect on both theory and practice.

Reflective Learning Cycle[5]



A praxicum must be distinguished from a practicum. The latter typically requires that a student engage in practical work under the supervision of a lawyer (or other professional). A praxicum, while incorporating the need to apply learned theory, also requires that students actively engage with and reflect on their experiences with a view—ideally—to allowing them to integrate theory and practice in an active, ongoing, and reflective way. The idea is to include elements of theory, practice and reflection—all as part of an engaged, continuous learning process. The purpose of such reflective education is to assist students to become reflective professionals (Farrow et al. 2012).

The Upper Year Curriculum Reform Working Group also presented a list of JD Program Learning Outcomes which are intended to extend beyond the baseline standard set out within Osgoode's UUDLEs (University Undergraduate Degree Level Expectations), which are required of all faculties within York University. The Program Learning Objectives, consistent with the recommendations of the *Carnegie Report*, focus on Knowledge, Skills, and Values.

7.5 Osgoode's JD Program Learning Outcomes

The overarching objective of Osgoode Hall law School's JD Program is to integrate knowledge, skills and values in the development of reflective legal professionals.

[5] Farrow et al. (2012).

Specific elements of this objective are set out in the following list of key learning outcomes. This list is not exhaustive and will be continuously reviewed and renewed.

A JD graduate from Osgoode is expected to:

Knowledge

1. Understand and apply the principles, rules, procedures, and theories of law in a variety of institutional settings.
2. Analyze and situate the law in its political, social, economic, and historical context.
3. Understand and apply various approaches to dispute resolution.

Skills

4. Conduct legal research at an advanced level in NGO, policy, law firm, or academic settings.
5. Integrate knowledge, skills, and values; theory and practice; some dimensions of a discipline other than law; international and domestic law; and different cultures, systems, and ideas.
6. Act in an ethical and professional manner acknowledging the complexities of different perspectives and sensibilities in a global community.
7. Analyze legal problems creatively, both individually and in collaborative settings.
8. Communicate effectively in written and oral contexts.

Values

9. Exhibit an awareness of self in the context of a pluralistic community.
10. Critically reflect on their role in society and on the limits of the law demonstrating the abilities of flexible and adaptive thinking.
11. Assume responsibility for serving and promoting social justice and act in the interest of the public good.

As demonstrated by the Program Learning Objectives, Osgoode is institutionally committed to creating opportunities to ensure that students graduate with the combination of knowledge, skills, and values described above. Osgoode is also cognizant of the Federation of Law Societies competencies requirements. It has developed two online modules—"Fiduciary Relationships in a Commercial Context" and "Principles of Administrative Law Lecture" which students must complete if they have not taken the Business Associations course and the Administrative Law course, respectively. Completion of either the relevant course or online module ensures that Osgoode students have been exposed to these topics to the extent required to meet the Federation's statement of required competencies in these areas.

7.5.1 Course Offerings

In an effort to illustrate the range of teaching approaches encompassed within the Osgoode curriculum, the seminar, course, intensive and clinical offerings were analyzed using Katz' categorization of major educational settings (that do not fall within the traditional doctrinal teaching parameters):

- Clinics or externships
- Skills-focused simulation courses
- Practice context simulation courses
- Doctrinal courses with skills exercises
- Doctrinal courses that critically analyze lawyering

A review of the 2013–2014 upper year offerings (consisting of 160 courses, seminars, or intensive/clinical programs) show 15 clinical and intensive programs, with 189 students participating. Sixteen courses fell within the skills-focused simulation category with 333 student enrolments, and 15 fell within the practice context simulation courses (note that some of these also contained skills-focused simulation exercises. However, the courses have been counted in only one category), with a total of 147 enrollments. For upper year students, these two categories provide the greatest number of opportunities for "nontraditional" learning approaches. Seven doctrinal courses offered skills exercises. The enrollment in these courses during the 2013–2014 academic year was 60. Finally, 12 courses were categorized as doctrinal courses that critically analyze lawyering. Seventy students enrolled in these courses.

In the first year curriculum, all 289 students were required to enroll in both Legal Process and Ethical Lawyering (described above). Legal Process falls both within the skills-focused simulation and practice context simulation course categories, while Ethical Lawyering offers both practice context simulation exercises and critical analysis of lawyering.

In summary, all first year students are required to enroll in two courses which combined offer skills-focused simulation, practice context simulation, and critical analysis of lawyering. Within the upper year curriculum, 787 spots were filled within courses offering one or more type of learning approaches described by Katz. There is obviously overlap occurring, so that some students will take more than one of these upper year courses while conceivably some students may have enrolled in none of these courses. However, as a result of the praxicum requirement adopted by the faculty within the Upper Year Curriculum Reform initiative, students in the graduating class of 2015 and beyond will have been required to take at least one course that complies with the praxicum requirement. Students do still have a significant amount of choice in terms of the way that they craft their learning curriculum throughout their time at law school. While the first year curriculum is mostly mandatory, second and third year courses are left to student selection.

For the purposes of the current work, interview data from four instructors are analyzed. The data relate to the following course, seminar, or program: Advanced Business Law Workshop: Corporate Finance; Theory and Practice of Mediation; Criminal Intensive Program; and Collaborative Lawyering. The instructors employ a range of learning objectives and assessment methods.[6]

7.5.2 Analysis of Syllabi and Interviews

Osgoode has a robust selection of courses, seminars, clinical placements, and intensive programs. The syllabus for each of the programs, courses and seminars described above, and the four instructor interviews were analyzed with reference to Mentkowski's outcomes. Each element is discussed below, with selected representative passages illustrating the manner in which the elements appear in either the learning objectives, description of evaluation for the particular program, course or seminar within the syllabus, or instructor interviews.

Self-reflection. There is a great deal of emphasis on self-reflection within the syllabi analyzed. For example, the trial advocacy course, which had an enrollment of 89 students during the 2013–2014 academic year, provides that one of the course objectives is to allow students to "develop the capacity to engage in critical self-reflection regarding their own professional role and performance."

Reflection is included as an evaluative component by way of class discussions, maintenance of reflective journals, and the creation of research papers containing reflective components. One intensive program—the Intellectual Property Intensive—requires an internship reflective journal, and an internship reflective blog. Another—Lawyer as Negotiator—requires that students submit a midterm reflection, which discusses their learnings from their own midterm negotiation experience.

Reflection was also discussed in some detail within the instructor interviews. Each interview participant interviewed discussed the role that self-reflection played within their program. Reflection formats spanned reflective journals, papers incorporating reflective components, verbal reflection at the completion of a role play or simulation, short written "learning points" provided at the end of each class, and class discussion. Instructors reported on students grappling with course concepts and experiences in ways that linked to their individual growth, their collective efforts, and their experience of the legal system:

Grappling with Course Concepts: [from an instructor who had used reflective journals in the past]

- I would have a better sense when I was reading the journals from week to week of what the class was struggling with, what they were questioning, what they were not getting…

[6] A copy of the course syllabi for these course offerings is on file with Shelley Kierstead.

Individual Growth: [from the same instructor commenting on the paper assignment, which required the incorporation of personal reflection as a paper component]

- The papers were good in that they did, I would say, more than half of them would put something in there about how they changed their thinking from the beginning to the end of the course, which is pretty typical for a student's experience of our course. Many students come in very cynical or really questioning what this crazy thing called 'collaborative' is, and really fighting it in the first couple of classes. That's also what we see in our adult training about all this stuff too, so that's part of what might be called the paradigm shift. Part of it is sort of feeling your resistance and working through it. A lot of the students in their papers did talk about that, "Initially I was of the view that XYZ, but by the end of the course I really started to wonder if ABC."

[From another instructor who discussed student reflections after watching role plays that had been video recorded]:

- "I can't believe … that much learning could come out of this experience. I had one student in … [a] class last term who was so quiet. Whenever he spoke it was the most poignant perfect reflection. Clearly he was taking it all in and thinking exactly how this worked in his life."

[And, from an instructor describing a final "wrap up" class discussion] …

- [W]e spend another hour going through the course. What did you learn? How did you learn it? What did you find was useful? Think about what you learned on a particular assignment and what was it about that assignment that helped you learn that? I guess that's kind of that aspect, that one hour when we think back through the course. What did you learn and how did you learn that is a bit reflective.

The Legal System: [from an instructor who commented on the required submission of learning points at the end of each class noted above]

- It has a bit of a critical self-reflection component where a lot of them have some pretty thoughtful things that they say and they're clearly thinking about the subject matter that they've dealt with or the group discussion. … It's more of a critical reflection to look back at your week, how it was, what you spent time doing, what you found interesting in the system, shocking, disturbing, all that stuff.

Overall, the analysis of syllabi and interviews suggests that the curriculum provides rich opportunities for student reflection on their own performance as individuals about to enter the legal profession, and on the challenges that the legal profession might pose for them.

7.5.2.1 Learning Metacognitive Performance

The focus on fostering the development of metacognitive performance is evident within some of the course syllabi analyzed. It is most evident within the materials for the intensive and clinical programs. One intensive program—community action—describes its aim as follows:

> A central aim of this course is for students to learn how to make a difference: how to be a community-based lawyer on a global scale. It offers students the opportunity for skills training, hands – on experience, structured reflection and peer collaboration. It merges theory, doctrine and practice in a dynamic, comprehensive and multi – disciplinary setting.

A seminar program titled CLASP case seminar is required of students who participate in the student legal aid clinic. This seminar specifically aims to critically assess fundamental skills such as ethical lawyering, reflective lawyering, pluralism, client communication, and to adopt a harmonized approach to problem solving using legal and nonlegal options. Students within this seminar are expected to draw on the breadth of their clinical experience to provide opportunities within class discussions and their written work to undertake a comparative analysis in different substantial legal settings.

A legal drafting course aims to foster metacognition in the context of document formation: "… [W]e are not focusing just on the acquisition of skills—our aim is to enhance our understanding of the way documents shape and form our lives—and how we as lawyers contribute to that process."

7.5.2.2 Self-assessing Performance and Using Feedback to Transform It over Time

Likewise, there is evidence of the use of self-assessment of performance and use of feedback to transform performance over time. Different tools are designed to foster this transformation. For example in the trial advocacy program students are videotaped every third session. After performing the particular skill at issue, and being critiqued by two faculty members, students will review their videotaped performance with another instructor. The repetition of the videotaping and debriefing of the videotape is designed to allow students to improve their performance over time. Another example is illustrated by the disability law intensive which provides that:

> Students will be evaluated informally and formally by the staff lawyers at ARCH [Advocacy Resource Centre for the Handicapped]. Informally, students will receive feedback on their work through discussions with supervising lawyers and written comments and requests for revision on written work. In addition, students will receive a formal evaluation from their supervising staff lawyer [using a structured evaluation form that will be made available to students] halfway through each semester.

This feedback mechanism is designed to provide students with the ability to transform their performance over time. The assessed mediation within the mediation intensive program is aimed at a similar objective:

> All students will participate in an assessed mediation. Students will be evaluated on the development of both their mediation and conflict resolution skills. The mediation will be similar to mediations undertaken throughout the semester and in the seminar and workshop simulations. Students will be asked to reflect on their roles as mediators and participate in a self-critique following the mediation.

During an interview, the instructor for the mediation intensive program commented on the process of having students receive feedback and reflect on their early mediations:

- "While two cases are never going to be the same, now they could be vastly different. At least they do have that, okay, well this time this happened. For next time I'm going to change this or that, or look for a certain thing, or whatever."

7.5.2.3 Professional Expertise

The development of professional expertise is an objective that a number of programs seek to foster through their performance assessments. For example, within the lawyer as negotiator course, students are required to submit negotiation plans prior to conducting actual negotiations. These plans must provide a strategic analysis of the issues and potential negotiation strategies for the particular negotiation. In the trial advocacy program, students are required to conduct a trial from beginning to end "with reasonable competence" after having worked on the discrete elements of a trial throughout the term. And in the Child Protection course described in detail below, students conduct a temporary care hearing during class and prepare a short reflective paper about this experience as a child protection advocate. Students who enroll in the Constitutional Litigation course draft constitutional pleadings and a brief of materials to be proved within an assigned case.

While the four courses profiled in the interview portion of the case study focused on different aspects of professional expertise, some commonalities emerged. Each course contained significant expectations that students perform across the four elements of communication (influencing and advocating, writing, speaking, and listening). In particular, students were expected to pay attention to grammar and production, which some found surprising. Speaking and listening were significant features in the classroom, as well as in actual or simulated performance. Influencing and advocating were demonstrated through familiar performances, such as trials, as well as less familiar, such as advising a client. Conflict resolution skills were a particular focus for the mediation and the collaborative lawyering course.

Students were expected to undertake research and information gathering, at least within the context of course papers and within their criminal intensive placements. While every course included an intellectual and cognitive component contained in a written document (research memorandum, trial document, Ontario Securities Commission comment letter, etc.), the focus was different across the courses. The two substantive courses, advanced business law and criminal intensive, required that students undertake traditional legal analysis of statutes, cases, and regulatory documents. The mediation and collaborative lawyering course instructors expected students to analyze the theory and practice of the lawyering experience, together with reflection on when and how the theory worked in practice.

Two elements of professional expertise were less prominent: client and business relations (apart from the business law course), and the “evaluation, development, and mentoring element of Working with Others” (as described within the Zedeck and Schultz framework). Finally, it was clear that the framework elements were assessed synthetically rather than analytically. Each performance contained multiple elements, generally including legal analysis, problem solving and communication.

- “Sometimes it’s not purely just their understanding and comprehension of the matter, sometimes it’s their ability to present and articulate well.”
- “We ask questions just as if we were like judges, and see how they respond to them not only in terms of their knowledge, but in terms of their character.”

7.5.2.4 Uncertainty in Performance

While not asked directly, participants did refer to making decisions in conditions of uncertainty

- “Some of the answers in the memo will be right or wrong. They need to know if they require this many shares, they’ve crossed the takeover threshold…but a great deal of it is not substantive. It is applying the knowledge of the law to the business issue at hand and coming up with advice and recommendations as to strategy, and pros and cons of various strategies.”
- “I would never send a student to small claims court without a coach. It’s just too risky. I think they’d be fine, but they need that. There was once a coach who didn’t debrief an unethical dilemma a student faced and the student was quite upset about it for a long time afterwards.”

7.5.2.5 Identity

There is evidence of the focus on developing one’s identity as a professional and a unique learner within the course syllabi. Many references to this development are found within the stated objectives of the particular curricular offering. For example, the Parkdale Poverty Clinic syllabus states that to “identify and assess the interplay of professional and personal boundaries” is one of the objectives for students who work within the clinic. In another intensive program—Criminal Law Intensive—the course outline indicates that one of the course objectives is to allow students to develop the capacity to “engage in critical self-reflection regarding one’s own professional role and performance.”

The multidimensional approach to teaching and assessment described by the instructors interviewed for this study can be linked to the goal of assisting students to develop their sense of professional identity. In particular, instructors commented on lectures, role plays, simulations, and case studies.

In one course, **lectures** are often supplemented by **guest speakers** who are practitioners. The practitioners give students a sense of what they are likely to be exposed to within their placements:

- Initially, there's a two week component to it where they're in class every single day. We're giving lectures and we have guest speakers that come in. … It's really to get their feet wet to get a sense of what they're going to be up against when they go to their placements …

Role plays, simulations and case studies allow students to experience the mix of substantive, procedural and ethical problem solving involved in legal practice:

- A lot of the learning has to be, in this kind of course, the learning really isn't the lecture style and that passive receiving of information really only allows you to answer on paper. The real life examples and then role playing on the spot, if somebody had a question about, "Well if I were a client I would do this." I could immediately invite them, "Okay, you want to be the client and I'll be the lawyer in that initial interview and let's role play it right here." You can put your concerns right into a quick mini role play and stuff like that.

Student reflections, either verbal or written, about these experiences, as reviewed above, allow and require students to think deeply about how they as individuals will take up these challenges. Finally, the **feedback** that they receive on both the experiences and their reflection on those experiences provides students with additional information with which to continue to formulate their sense of professional identity:

- "I kind of felt like that immediate feedback, the in-the-moment feedback right after [role plays] was probably where some of the biggest learning of the year happened."
- "They [practitioner coaches] then meet with the students afterward (after mediation role plays) and debrief with them. … Give some suggestions for the future and hopefully at that point it'll elicit any issues the students might have felt in terms of ethical issues of questions the students might have.
- "At the end of the year, the placements give us an evaluation form in terms of how the student did in their placement. What kind of duties they performed, and how were their legal research skills, their personable skills, anything else that the placements want to add in terms of evaluation. They typically will give some pretty fair and honest opinions. They typically end with what kind of lawyer they think they're going to be."

The syllabi reviewed and the instructors interviewed demonstrate a significant focus on creating a rich learning experience that provides students with methods of evaluation intended to match the learning objectives set out within the course syllabi. Neither the overview of the course syllabi nor the interviews provide an in-depth look at the actual evaluation instruments. This is the next logical step in this research. It appears, however, safe to say that more focus has been placed on the development of learning objectives and varied forms of feedback then on an in-depth evaluation of the effectiveness of the assessment tools.

7.5.2.6 Teacher Commitment

Our original interview protocol did not explore aspects of teacher commitment as a feature of performance assessment. However, in the initial interview we conducted (not included in this study) the professor noted that he dedicated at least a day for each student to provide feedback on their work. As we incorporated this question into subsequent interviews, we found that our participants also devoted many hours to student feedback:

- …when we hand back the credit agreements, they take so long to mark…We don't just hand them back with a mark, right? We make all sorts of comments on them and so forth, because otherwise, I don't think it's much of a learning experience.
- Then we provide specific feedback for each student on the evaluation form on how we think they did on the seminar, giving some comments there. How they did on both papers, just to flush out more than just the letter grade…feedback in terms of the grading they got for the papers. We give them some feedback on the advocacy component to it. We give feedback from the placements themselves, and incorporate anything that we think should be incorporated in that so they're got a pretty substantial evaluation form of probably seven or eight pages long that they are free to use.

Participants also discussed methods to improve their courses, either during the term or at the end. One participant required students to submit a short "learning point" at the end of each class. Others conducted a full debrief at the end of term to reflect on learning throughout the course.

In the final section of this case study we describe a course offered by one of the authors, and offer initial comments about next steps for performance assessment.

7.5.3 Child Protection Course Example

The Child Protection course described below was offered during the fall terms of 2011 and 2013. On the latter occasion, it was a prerequisite for students who participated in a Child Protection externship program during the Winter 2014 term. The externship allowed students to attend the offices of various professionals who are involved in child protection work. The students were also asked from time to time to assist in research or file review of the matters that their professional mentors were dealing with when the students attend at their offices.

The objectives for the Child Protection course were designed based on four overarching themes: (1) the need to understand children's special status and unique relationship with the law; (2) the importance of understanding socioeconomic factors that impact the dynamics of state—family interaction in child welfare cases; (3) the importance of core lawyering skills such as critical thinking, effective speaking and writing, and solid legal research to this specific area of law; and (4) the uniquely interdisciplinary nature of this area. The learning objectives are reproduced below.

7.5.3.1 Course Objectives

- Gain a solid understanding of the evolution of our understanding of "childhood" and how this evolution impacts parental and state relationships with children
- Identify the ways in which law can affect children's lives
- Recognize the institutions and people who exert power or influence over children and families
- Recognize societal conditions that impact families' ability to operate independently from state intrusion
- Identify ways in which law and policy can protect children from abuses of power
- Evaluate the effectiveness of domestic legislation designed to protect children
- Develop strategic case preparation and evaluation skills
- Develop an understanding of the interdisciplinary nature of child protection matters
- Demonstrate an ability to think critically and to justify ideas in a reasoned manner
- Communicate effectively in both speaking and writing
- Conduct advanced level legal research and writing
- Gain insights from respectful dialog among fellow students and instructors who may consider topics from various perspectives/positions

Course Evaluation. The course evaluation involved a number of components:

(i) Required Exam (50 %)
Students wrote an open book examination consisting of a fact pattern which required them to identify issues, consider appropriate statutory and case sources to respond to the issues, and assess practical strategic considerations that ought to be brought into play in resolving the matter. The exam also contained a number of short answer questions designed to allow students to demonstrate their understanding of the domestic law that governs child protection matters.

(ii) Role Play and Reflection (25 %)
Students conducted a temporary care hearing—that is, a mock court process that occurs shortly after a child is apprehended from his or her parent/guardian's care. The hearing is widely recognized as playing a key role in shaping the course of the remainder of the process. Students were assessed on both their substantive understanding of the case and on their reflection after the completion of the exercise. The students were given the following instructions with respect to this exercise:

Role Play and Reflection Evaluation Guidelines—October 2013.

1. Role Play
During class time on October 22, you will act as counsel for either the Children's Aid Society or parent (as assigned in class) with respect to Case

Study #2. This case study has been posted on the course site. Each student will have 20 min to present his or her submissions.

The role play will be graded out of 10 possible marks. These marks will be allocated as follows:

- Student has formulated an appropriate "theory of the motion." **(2.5 marks)**
- Student understands which facts (positive and negative) are most significant to the outcome of the hearing, and addresses them appropriately in relation to his/her client's position. **(2.5 marks)**
- Student demonstrates an understanding of the statutory provisions that govern this temporary care hearing. **(2.5 marks)**
- Student is able to discuss how the facts of this case relate to the governing statutory provisions. **(2.5 marks)**

2. Written Reflection

Following the temporary care hearing, you will be asked to submit (by October 29, 2013 at 4:30 p.m.) a short reflection paper (maximum 1000 words) about your participation in the temporary care hearing. The reflection will be graded out of 15 possible marks. Of these, 10 marks will be allocated to evidence of your thoughtful consideration of the following issues, along with any others that you deem significant:

- What was your personal comfort level with the role you assumed? Did it fit with your overall understanding of "fearless advocacy"?
- How, if at all, did your view of the strength of your client's position (Society counsel's "client" is the instructing social worker) impact the arguments you made?
- Would your approach to the arguments change if your view of the strength of your client's position was the opposite of that which you discussed above?

The additional 5 marks will be allocated to proper grammar, sentence structure, and coherence of presentation.

(iii) Case Study Preparation (25 %)

In the middle of the term students were given a fact pattern which was discussed briefly during class. They were responsible for preparing a 1500 word memo identifying key issues arising from the file, the potential interests of the various parties described in the fact pattern, and potential avenues for moving toward legal resolution of the identified issues. The fact pattern raised difficult questions without clear legal solutions, and was set within a backdrop that required consideration of underlying societal conditions and their potential impact on the parties. It was also designed in a manner that required students to incorporate feedback from the temporary care hearing role play. The grade was broken down into the following grading categories: accurate and complete identification of issues: 5 %-assessment of parties' (competing and overlapping) interests; 5 %-identification and application of legal principles; 5 %-consideration of both legal and nonlegal strategic considerations; and 5 %; proper grammar, sentence structure and style.

Assessment methods were designed to reflect the themes that the learning objectives encompassed. Table 7.1 provides an example.

Reflection on Assessment Methods. While some aspects of the assessments addressed slightly different aspects of the course learning objectives, they were all aimed at helping students to develop their ability to make professional decisions in conditions of complexity and uncertainty. The role play and reflection were also specifically designed to have students consider their own role in resolving disputes involving important rights of parents and children. The extent to which students grappled with these issues was impressive:

> One of the challenges of representing Ms. Smith was my awareness that I was constructing arguments with real life implications without having an experience as a parent or with alcoholism. This experience reminded me of articles in the course materials that refer to how professionals (lawyers and social workers especially) predominantly come from different backgrounds to individuals participating in child protection proceedings in the readings.
>
> …
>
> One issue I faced was, if my arguments were successful, and the child was returned to her care, would I be able to say that I acted in the child's best interest? Is this even my role? Am I allowed to act in the best interest of my client at the expense of the child's best interest? This was a part of the balancing that was necessary because I was not sure if I should be acting in the best interest of the client, in light of the child's best interests, or if I should solely be acting in the best interest of the client that I'm representing.

Overall, the combination of assessments, and the students' performance on the assessments, provided a level of confidence that they had "matched" the learning objectives set out for the course. Further, there was evidence of students having had the opportunity to reflect on their learning abilities and use feedback to transform their performance. Anecdotally, based on group debriefing sessions relating to the externship that followed the 2013 class, students reported a solid understanding of the "real life" practice issues to which they were exposed, suggesting the successful

Table 7.1 Example of course activities and assessment methods

Course themes (from which learning objectives were derived)	Associated activity and performance assessment method
Children's special status and unique relationship with the law	Role play/reflection
Socioeconomic factors that impact the dynamics of state—family interaction in child welfare cases	Case study; role play/reflection
Importance of core lawyering skills such as critical thinking, effective speaking and writing, and solid legal research	Role play; case study

development of professional expertise. Student reflections also suggested the beginnings of a sense of identity flowing from the experiential exercises. As the course evolves, additional changes are planned: (1) Provide a more detailed rubric for the case study memorandum; and (2) video-tape role plays and de-brief them with students, so that they can understand exactly where the feedback on their oral advocacy performance applies. This will potentially allow for additional depth of self-assessment and self-reflection.

7.6 Discussion and Conclusion

Our review of the literature on performance assessment in law schools generally, and our specific analysis of the Osgoode Hall Law School curriculum, confirms that a great deal of curricular innovation and corresponding assessment development is occurring. Further, a high level systems approach is evident within developments such as the Praxicum requirement. Students have ample opportunities to learn a full range of skills and values. The combination of required first year courses that incorporate a number of skills areas and their ability to choose from a wide range of curricular offerings during their second and third years result in students having the ability to create an individualized program of assessment. The scope of this program is largely within their control.

The examination of course syllabi, analysis of instructor interviews, and in-depth look at the child protection course at Osgoode Hall Law School suggest that faculty members are designing their curricular offerings in ways that foster a number of the Mentkowski outcomes.

One obvious element arising from the analysis of course syllabi and interviews is the faculty focus on having students employ self-reflection methods (such as journals, papers, oral reflection) in order to engage deeply with their own learning abilities. Likewise, a number of the courses analyzed provide feedback to students so that they may use it to transform their performance over time (for example, engaging in mediations for which feedback and reflection are required, and incorporating that learning in preparation for the final graded mediation). The development of professional expertise is perhaps the most clearly demonstrated element of performance assessment evident from the case study. Externship placements, role plays and simulations, coupled with feedback and reflection, provide rich opportunities for students to develop professional expertise. These offerings reinforce the potential for students to experience transformative learning cycles.

While there is some evidence within course syllabi of instructors' intention to assist students to develop metacognitive performance, there was relatively little discussion of this element of performance assessment within the interviews. We may infer that the transformation in learning that students experience as a result of their participation in the praxicum courses studied would assist them to engage in restructuring their knowledge. However, more nuanced questions, and perhaps interviews with students, would provide a clearer picture of this potential outcome. Likewise, while there is evidence of students' development of identity (for example, student reflection on his/her role in navigating the complex mix of interests and sociodemographic realities involved in child protection matters), further research is required to obtain a more fulsome sense of the extent to which this element is reflected in current course offerings. Next steps for law schools could include developing an understanding of how each of the courses surveyed addresses the role of legal reasoning, problem solving and decision-making under conditions of uncertainty. Further, it would be helpful for institutions to undertake an examination of the range of methods used across their curricula to assess different competencies in different settings.

Institutionally, law schools should support movement toward clarity in the creation, analysis and reporting of performance assessments, including psychometric and/or qualitative properties. In this regard, much could be accomplished through a collapse of "silos" in legal education, that is, the division of law school faculty into doctrinal, clinical, and legal writing/process teachers and researchers. Traditionally, legal education has been divided between doctrinal professors and others (clinical educators, skills teachers, and legal research and writing instructors), with the "others" enjoying less prestige within the law school hierarchy (Glesner Fines 2013, p. 21).

Courses may be taught by specialized faculty such as clinical legal educators, practitioners, regular faculty, or in the case of externships, knowledgeable principals at the externship placement. Courses may differ substantially in their philosophical mission, which affects their approach to pedagogy and assessment. Clinicians, for example, have developed sophisticated approaches to reflective practicum, while legal writing and research academics focus on detailed approaches to feedback and continuous improvement. Clinical educators and legal writing instructors tend to publish in their specialized journals (Clinical Law Review, Journal of the Legal Writing Institute, Journal of Legal Writing and Rhetoric, International Journal of Clinical Law) and attend their own conferences—so that their approaches and academic knowledge on performance assessment is not shared or understood across the law school setting.

However, each academy has much to offer in the development of sophisticated performance assessments. For example, legal writing instructors have finely tuned rubrics, with a focus on legal analysis, problem solving, and communication. Skills courses can provide strategies for self-assessment through their focus on feedback on professional expertise. Clinics and externship programs have highly developed reflective course components, often including analyses of professional identity and the role of lawyers in society. Finally, doctrinal courses contribute to students' development of problem solving and analysis in specific contexts. Garth and Martin (1993) argue that law schools not only develop competence; they also construct competence. Law schools have an opportunity to enhance the construction of competence through the rigorous reporting and sharing of assessments. Student insight into their growth and development as practitioners—as evidenced by the brief review of the Child Protection course—illustrates one such construction of competence. Their broad range of curricular offerings makes law schools the ideal venue for analysis of the theory-practice–reflection cycle. Future investigation is required to explore the extent to which students' personal transformation results in a collective shift within the profession—toward an integrated vision of the lawyer.

Issues/Questions for Reflection

- Should law schools keep records of the extent to which students engage in a broad array of simulation/performance courses?
- Should law schools develop portfolio assessments that are linked to entry-level competencies?
- How best can educational researchers focus on the extent to which simulation courses support students as they enter practice?
- Future investigation is required to explore the extent to which students' personal transformation results in a collective shift within the profession —toward an integrated vision of the lawyer

Appendix 1: The Continuum of Legal Education in the United States and Canada

The continuum of legal education in both Canada and the United States results in a common end point: admission to a particular jurisdiction as a member of a fused (that is, both barrister and solicitor) profession with a general license.

United States*

As a general overview, the path to bar admission includes a three-year law degree, passing the state-administered Bar Admission Examinations, and providing evidence of sound character and fitness. About a third of jurisdictions restrict eligibility for licensure to graduates of American Bar Association (ABA)-approved law schools. Other jurisdictions allow graduates of non-ABA approved law schools to apply for licensure, but typically only under certain conditions (such as already being admitted by examination in another jurisdiction and/or meeting active practice requirements). An overview of the process can be found at http://www.americanbar.org/groups/legal_education/resources/bar_admissions.html and for the bar admission examinations at the website of the National Conference of Bar Examiners (NCBE) http://www.ncbex.org/. NCBE prepares and coordinates administration of four national exams: the Multistate Bar Examination (MBE), the Multistate Essay Examination (MEE), the Multistate Performance Test (MPT), and the Multistate Professional Responsibility Examination (MPRE). Jurisdictions differ in terms of which exams or components of exams they require, with some requiring jurisdiction-specific exams in addition to or instead of NCBE-produced exams. Over the past few years many jurisdictions have adopted the Uniform Bar

Exam (UBE), which combines MBE, MEE, and MPT scores in a standardized way and allows for greater score portability across jurisdictions. For more information about the individual exam components and the UBE, see Chap. 20 in this volume.

Following admission, each state determines continuing legal education requirements, which may be more stringent in the early years of practice. Performance in practice is monitored through the human resources/performance management practices at each setting. In this regard, it is important to note the growth of in-house professional development practitioners, dedicated to associate learning and development. More information on this topic can be found at the website at the National Association of Legal Career Professionals [NALP]: http://www.nalp.org/.

Canada

In contrast to the United States process, the path to bar admission includes a year of apprenticeship and, with the exception of Ontario, bar admission examinations that include performance assessments. Law schools are accredited by the Federation of Law Societies of Canada; there are no unaccredited law schools. Lawyers are regulated through a provincial law society, which sets out the path to practice including Bar Admission process and licensing examinations. Bar Admission courses are developed and administered within each province by the provincial regulator; this oversight includes the performance assessments and other examinations. Bar Admission courses may include an up to three-month in-class program taught by practitioners. Mandatory continuing legal education is also provincially determined. The Federation of Law Societies of Canada 2015 introduced a National Admission Standards Project, aimed at developing and implementing a national assessment regime.

Further overview information can be found at: http://www.flsc.ca/en/national-requirement-for-approving-canadian-common-law-degree-programs/.

While there is modest regulation by the provincial law society of the articling year, there is no further regulation after bar admission. Canadian professional development practitioners are also involved in NALP and have also formed a Canadian section. *Prepared by Joanne Kane, Associate Director of Testing, National Conference of Bar Examiners

Appendix 2: Competency Frameworks

MacCrate Report (American Bar Association Section of Legal Education and Admissions to the Bar 1992) contains extensive discussion of the content of each of the following skills and values.

Fundamental Lawyer Skills

1. Problem Solving
2. Legal Analysis and Reasoning
3. Legal Research
4. Factual Investigation
5. Communication

6. Counseling
7. Negotiation
8. Litigation and Alternative Dispute Resolution Procedures
9. Organization and Management of Legal Work
10. Recognizing and Resolving Ethical Dilemmas

Fundamental Values of the Profession

1. Provision of Competent Representation
2. Striving to Promote Justice, Fairness and Morality
3. Striving to Improve the Profession
4. Professional Self-Development

Schultz and Zedeck (2011, p. 630) **List of 26 Effectiveness Factors**

1. Intellectual and Cognitive
   - (a) Analysis and Reasoning
   - (b) Creativity/Innovation
   - (c) Problem Solving
   - (d) Practical Judgement
2. Research and Information Gathering
   - (a) Researching the Law
   - (b) Fact Finding
   - (c) Questioning and Interviewing
3. Communications
   - (a) Influencing and Advocating
   - (b) Writing
   - (c) Speaking
   - (d) Listening
4. Planning and Organizing
   - (a) Strategic Planning
   - (b) Organizing and Managing One's Own Work
   - (c) Organizing and Managing Others (Staff/Colleagues)
5. Conflict Resolution
   - (a) Negotiation Skills
   - (b) Able to See the World Through the Eyes of Others
6. Client and Business Relations—Entrepreneurship
   - (a) Networking and Business Development
   - (b) Providing Advice and Counsel and Building Relationships with Clients

7. Working with Others
   (a) Developing Relationships within the Legal Profession
   (b) Evaluation, Development, and Mentoring
8. Character
   (a) Passion and Engagement
   (b) Diligence
   (c) Integrity/Honesty
   (d) Stress Management
   (e) Community Involvement and Service
   (f) Self-Development

Law Firms

Law firm frameworks can be difficult to find, since many are considered proprietary. However, organizations such as the National Association for Law Placement have included sessions on competencies in their annual conferences, where a number of firms have made presentations. Some published literature includes detail on these frameworks.

Sloan's "Associate Level System" (2002) was developed entirely within a single law firm, finding 17 competency areas and articulating specific expectations for admission to partnership. These competency areas were divided into four levels. For example, the competency of Written Communication is described as follows:

Level 1: Drafts clear and concise correspondence, pleadings, legal memoranda, or transactional documents, for review by supervising lawyer.
Level 2: Drafts clear and concise correspondence, pleadings, legal memoranda, or transactional documents, for review by supervising lawyer and requiring few modifications.
Level 3: Takes primary responsibility for most correspondence, pleadings, legal memoranda, or transactional documents, with minimal review by supervising lawyer. Work product is clear and concise.
Level 4: Takes primary responsibility for correspondence, pleadings, legal memoranda, or transactional documents, with supervisory responsibility over other lawyers working on less complex matters. Work product is clear and concise (Sloan 2002, p. 21).

References

AALS Conference on Clinical Legal Education. (2010). Paper presented at the Answering the Call for Reform: Using Outcomes Assessment, Critical Theory and Strategic Thinking to Implement Change.

Alford, D. (2009). The Development of the skills curriculum in law schools: Lessons for directors of academic law libraries. *Legal Reference Services Quarterly, 28*(3–4), 301–319.

American Bar Association Section of Legal Education and Admissions to the Bar. (1992). Report of the task force on law schools and the profession: Narrowing the gap. Legal education and professional development—An educational continuum (The MacCrate Report). Chicago: American Bar Association.

Backman, J. H. (2006). Where do externships fit? A new paradigm is needed: Marshaling law school resources to provide an externship for every student. *Journal of Legal Education, 56*(4), 615–655.

Backman, J. (2013). Significant but unheralded growth of large externship programs. Available on Social Science Research Network. http://ssrn.com/abstract=2235215

Barton, K., Cunningham, C. D., Jones, G. T., & Maharg, P. (2006). Valuing what clients think: Standardized clients and the assessment of communicative competence. *Clinical Law Review, 13*(1), 1–65.

Basu, S., & Ogilvy, J. (2012). Externship demographics across two decades with lessons for future surveys. *Clinical Law Review*, *19*(1).

Beazley, M. B. (1997). The Self-graded draft: Teaching students to revise using guided self-critique. *Legal Writing: Journal of the Legal Writing Institute, 3*, 175.

Berman, M., et al. (2014). Creative initiatives at U.S. law schools. *Second National Symposium on Experiential Education in Law: Experience the Future*, June, 2014.

Binder, D. A., & Bergman, P. (2003). Taking lawyering skills training seriously. *Clinical Law Review*, *10*(1).

Binder, D. A., Moore, A. J., & Bergman, P. (2006). A depositions course: Tackling the challenge of teaching for professional skills transfer. *Clinical Law Review, 13*, 871.

Bock, H. (2006). *Constructing core competencies: Using competency models to manage firm talent*: Amer Bar Association.

Brustin, S. L., & Chavkin, D. F. (1996). Testing the grades: Evaluating grading models in clinical legal education. *Clinical Law Review, 3*, 299–336.

Capulong, E. R., et al. (2015). *The new 1L: First-year lawyering with clients*. Durham: Carolina Academic Press.

Carpenter, C. L. (2012). Recent trends in law school curricula: Findings from the 2010 ABA curriculum survey. *The Bar Examiner*, *81*(2).

Case, S. (2013a). The NCBE job analysis: A study of the newly licensed lawyer. *The Bar Examiner*, *March 2013*, 52–56.

Case, S. (2013b). Summary of the job analysis survey results. www.ncbes.orga/publications/ncbe-job-analysis/

Cavazos, A. M. (2010). The journey toward excellence in clinical legal education: Developing, utilizing and evaluating methodologies for determining and assessing the effectiveness of student learning outcomes. *Sw. L. Rev., 40*, 1–57.

Chitwood, S., & Gottlieb, A. (2000). Teach your associates well: Developing a business and management skills curriculum for law firm associates. *Association for Legal Administrators*.

Clark, J., & DeSanctis, C. (2013). Toward a unified grading vocabulary: Using rubrics in legal writing courses. *Journal of Legal Education, 63*, 3.

Corrada, R. (2013). Formative assessment in doctrinal classes: Rethinking grade appeals. *Journal of Legal Education, 63*(2), 317–329.

Cort, R., & Sammons, J. L. (1980). The search for "good lawyering": A concept and model of lawyering competencies. *Cleveland State Law Review, 29*, 397.

Curcio, A. (2009). Assessing differently and using empirical studies to see if it makes a difference: Can law schools do it better? *Quinnipiac Law Review, 27*, 899–933.

Daicoff, S. S. (2012). Expanding the lawyer's toolkit of skills and competencies: Synthesizing leadership, professionalism, emotional intelligence, conflict resolution, and comprehensive law. *Santa Clara Law Review, 52*, 795.

Dijkstra, J., & Van der Vleuten, C. (2010). A new framework for designing programmes of assessment. *Advances in Health Science Education, 15*, 379–393.

Dinerstein, R., & Milstein, E. (2014). Learning to be a lawyer: Embracing indeterminacy and uncertainty. In S. Bryant, E. Milstein, A. Shalleck (Eds.), *Transforming the education of lawyers: The theory and practice of clinical pedagogy*. Carolina Academic Press.

Ebner, N., Cohen, J., & Honeyman, C. (2012). *Assessing our students, assessing ourselves*. St. Paul: DRI Press.

Engler, R. (2001). The maccrate report turns 10: Assessing its impact and identifying gaps we should seek to narrow. *Clinical Law Review, 8*, 109–169.

Eraut, M. (1994). *Developing professional knowledge and competence*. London: The Falmer Press.

Farrow, T., et al. (2012). *The praxicum requirement.* Osgoode Hall Law School, October 31, 2012.

Federation of Law Societies of Canada. (2012). National entry to practice competency profile for lawyers and Quebec notaries. http://www.flsc.ca/_documents/National-Requirement-ENG.pdf

Federation of Law Societies of Canada. (2015). National Admissions Standards Project. http://docs.flsc.ca/NASAssessmentProposalenSep32015(2).pdf

Fitzgerald, M. (1995). Competence revisited: A summary of research on lawyer competence. *Journal of Professional Legal Education, 13*(2), 227–280.

Friedland, S. (2013). The rhetoric of experiential legal education: Within the context of big context. *Northeastern University Law Journal, 6*(1), 253–286.

Gardner, H., & Shulman, L. S. (2005). The professions in America today: Crucial but fragile. *Daedalus, 134*(3), 13–18.

Garth, B., & Martin, J. (1993). Law schools and the construction of competence. *Journal of Legal Education, 43*, 469–509.

Gerkman, A., & Harman, E. (2015). *Ahead of the curve: Turning law students into lawyers.* Institute for the Advancement of the American Legal System.

Glesner Fines, B. (2013). Out of the shadows: What legal research instruction reveals about incorporating skills throughout the curriculum. *Journal of Dispute Resolution, 2013*, 159.

Grosberg, L. M. (2006). How should we assess interviewing and counseling skills. *International Journal of Clinical Legal Education, 9*, 57–71.

Grose, C. (2012). Beyond skills training, revisited: Spiraling the pyramid of clinical education. *William Mitchell Legal Studies Research Paper* (2012-11).

Grose, C. (2012–2013). Outcomes-based education one course at a time: My experiment with estates and trusts. *Journal of Legal Education, 62*.

Hertz, R., Bilek, M. L., Camper, D., Dennis, R., Dinerstein, R., Garth, B., et al. (2013). Twenty years after the MacCrate report: A review of the current state of the Legal Education Continuum and the Challenges Facing the Academy, Bar and Judiciary (pp. 1–24). Committee on the Professional Educational Continuum American Bar Association.

Hess, G., & Friedland, S. (1999). *Techniques for Teaching Law*. Durham: Carolina Academic Press.

Johnson, R., Penney, J., & Gordon, B. (2009). *Assessing performance: Designing, scoring and validating performance tasks*. New York: The Guilford Press.

Katz, H. (1997). Personal journals in law school externship programs: Improving pedagogy. *Thomas M. Cooley Journal of Practical and Clinical Law, 1*(7).

Katz, H. (2008). Evaluating the skills curriculum: Challenges and Opportunities for law schools. *Mercer Law Review, 59*, 909–939.

Kissam, P. (1989). Law school examinations. *Vanderbilt Law Review, 42*, 433.

Krieger, S. H. (2008). The effect of clinical education on law student reasoning: An empirical study. *William Mitchell Law Review, 35*, 359.

Krieger, S. H. (2011). Performance isn't everything: The importance of conceptual competence in outcome assessment of experiential learning. *Clinical Law Review, 19*, 251.

Law School Survey of Student Engagement. (2012). *Lessons from law students on legal education*. Indianapolis: Indiana University Centre for Post-Secondary Research.

Leering, M. (2014). Conceptualizing Reflective Practice for Legal Professionals. *Journal of Law & Society Policy, 23*.

Lopez, A. S., Crandall, C., Campos, G., & Rimple, D. (2009). Medical/Legal teaching and assessment collaboration on domestic violence: Assessment using standardized patients/standardized clients, A. *International Journal of Clinical Legal Education, 14*, 61–69.

Lynch, M. A. (2011). An evaluation of ten concerns about using outcomes in legal education. *William Mitchell Law Review, 38*(3), 976–1016.

Manch, S. G. (2010). Competencies and competency model—An overview. In T. Mottershead (Ed.), *The art and science of strategic talent management in law firms*. Thomson Reuters.

Martin, J., & Garth, B. G. (1994). Clinical education as a bridge between law school and practice: Mitigating the misery. *Clinical Law Review, 1*, 443–456.

Menkel-Meadow, C. (1986). Two contradictory criticisms of clinical education: Dilemmas and directions in lawyering education. *Antioch LJ, 4*, 287.

Mentkowski, M., Diez, M. E., Rauschenberger, M., & Abromeit, J. (2013, April). Conceptual elements for performance assessment for faculty and student learning: Paper prepared for the American Education Research Association.

Moss, D., & Curtis, D. M. (2012a). Essential elements for the reform of legal education. In D. Moss & D. M. Curtis (Eds.), *Reforming legal education* (pp. 217–230). Charlotte: Information Age Publishing.

Moss, D., & Curtis, D. M. (Eds.). (2012b). *Reforming legal education: Law schools at the crossroads*. Charlotte: Information Age Publishing.

Munro, G. (2000). *Outcomes assessment for law schools*. Spokane: Institute for Law School Teaching. http://lawteaching.org/publications/books/outcomesassessment/munro-gregory-outcomesassessment2000.pdf

Murray, V., & Nelson, T. (2009). Assessment-are grade descriptors the way forward? *International Journal of Clinical Legal Education, 14*, 48.

National Association for Legal Career Professionals, & The NALP Foundation for Law Career Research and Education. (2011). *2010 survey of law school experiential learning opportunities and benefits*. Washington: NALP.

Neumann, R., & Krieger, S. (2003). Empirical inquiry twenty five years after the lawyering process. *Clinical Law Review, 10*(1), 349–398.

Norcini, J., & Birch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide 31. *Medical Teacher, 29*(9–10), 855–871.

Norcini, J., Holmboe, E., & Hawkins, R. (2008). Evaluation challenges in the era of outcomes-based education. In E. Holmboe & R. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 1–9). Philadelphia, PA: Mosby/Elsevier.

Ogloff, J. P., Lyon, D. R., Douglas, K. S., & Rose, V. G. (2000). More than "learning to think like a lawyer": The empirical research on legal education. *Creighton Law Review, 34*, 73–243.

Professional Development Competence and Admissions Committee. (2004). *Report to convocation* (p. 83). Toronto, ON: Law Society of Upper Canada.

Ramy, H. (2013). Moving students from hearing and forgetting to doing and understanding: A manual for assessment in law school. *Capital University Law Review, 41*, 12–29.

Roper, C. (1999). *Foundations for continuing legal education*. Sydney NSW: Centre for Legal Education Law Foundation of New South Wales.

Sargent, C. S., & Curcio, A. A. (2012). Empirical evidence that formative assessments improve final exams. *Journal of Legal Education, 61*(3), 379–405.

Schön, D. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books Inc.

Schön, D. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Schön, D. (1995). Educating the reflective legal practitioner. *Clinical Law Review, 2*, 231–250.

Schwartz, M. H., Hess, G., & Sparrow, S. (2013). *What the best law teachers do*. Cambridge: Harvard University Press.

Sergienko, G. (2001). New modes of assessment. *San Diego Law Review, 38*, 463–506.

Sheppard, S. (1997). An informal history of how law schools evaluate students, with a predictable emphasis on law school final examinations. *UMKC Law Review, 65*(4), 657.

Shultz, M. M., & Zedeck, S. (2011). Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry, 36*(3), 620–661.

Sloan, P. B. (2002). *From classes to competencies, lockstep to levels*. Blackwell, Sanders, Peper, Martin LLP.

Sparrow, S. (2004). Describing the ball: Improve teaching by using rubrics-explicit grading criteria. *Michigan State Law Review, 2004*, 1.

Spencer, R. (2012). Holding up the mirror: A theoretical and practical analysis of the role of reflection in clinical legal education. *18 International Journal of Clinical Legal Education*, *17*, 181.

Stuckey, R. (2006). Teaching with purpose: Defining and achieving desired outcomes in clinical law courses. *Clinical Law Review, 13*(2), 807–838.

Stuckey, R. T. (2007). *Best practices for legal education: A vision and a road map*. Clinical Legal Education Association.

Sullivan, W., Colby, A., Wegner, J., Bond, L., & Schulman, L. (2007). *Educating lawyers: Preparation for the profession of law*. The Carnegie Foundation for the Advancement of Teaching: San Francisco, CA: Jossey-Bass.

ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of Graduate Medical Education, 5*(1), 157–158.

ten Cate, O., & Scheele, F. (2007). Viewpoint: Competency-based postgraduate training: Can We bridge the gap between theory and clinical practice? *Academic Medicine, 82*(6), 542–547.

Terry, K. (2009). Externships: A signature pedagogy for the apprenticeship of professional identity and purpose. *Journal of Legal Education, 59*, 240–268.

VanZandt, V. L. (2010). Creating assessment plans for introductory legal research and writing courses. *Legal Writing: Journal of Legal Writing Institute, 16*, 313.

Wegner, J. (2009). A legal education prospectus: law schools & emerging frontiers: Reframing legal education's "Wicked Problems". *Rutgers Law Review, 61*, 867–1127.

Wegner, J. (2013). Cornerstones, curb cuts, and legal education reform. *Journal of Display Resolution, 2013*(1), 33–84.

Wellford-Slocum, R. (2004). The law school student-faculty conference: Toward a transformative learning experience. *South Texas Law Review, 45*, 255.

Welsh, N. (2012). Making reputation salient: Using the reputation index with law students. In N. Ebner, J. Cohen, & C. Honeyman (Eds.), *Assessing our students, assessing ourselves* (Vol. 173–187). St. Pauls: DRI Press.

Ziegler, A. (1992). Developing a system of evaluation in clinical legal teaching. *Journal of Legal Education, 42*, 575–590.

Chapter 8
Assessing Student Learning Outcomes Across a Curriculum

Marcia Mentkowski, Jeana Abromeit, Heather Mernitz, Kelly Talley, Catherine Knuteson, William H. Rickards, Lois Kailhofer, Jill Haberman and Suzanne Mente

Abstract Disciplinary and professional competence in postsecondary education is made up of complex sets of constructs and role performances that differ markedly across the disciplines and professions. These often defy definition as learning outcomes because they are multidimensional and holistic. Even so, instructors who teach and assessors who evaluate competence in many fields may engage their colleagues in processes, usually within disciplines and professions, to capture enough breadth and depth of constructs and performances that are essential for particular roles. The question is whether students can integrate and transfer their learning across a curriculum and over time. Authors report on the design of an assessment technique for integration of knowledge constructs and role performances and their use, and adaptation and transfer across math and science prerequisite coursework. This assessment requires students to demonstrate scientific reasoning, quantitative literacy, analysis, and problem solving across these disciplines and over time, on demand, and in a setting outside of their regular coursework. During

M. Mentkowski (✉)
Faculty of Psycology, Alverno College, Milwaukee, WI, USA
e-mail: marcia.mentkowski@alverno.edu

J. Abromeit
Faculty of Sociology, Alverno College, Milwaukee, WI, USA

H. Mernitz
Faculty of Natural Sciences, Alverno College, Milwaukee, WI, USA

K. Talley · J. Haberman
Asseseement Center, Alverno College, Milwaukee, WI, USA

C. Knuteson
Faculty of Nursing, Alverno College, Milwaukee, WI, USA

W.H. Rickards
Research and Evaluation Department, Alverno College, Milwaukee, WI, USA

L. Kailhofer
Faculty of Mathematics, Alverno College, Milwaukee, WI, USA

S. Mente
Faculty of Instructional Services, Alverno College, Milwaukee, WI, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_8

training of faculty assessors, independent evaluators recorded and categorized faculty questions re validity and reliability of their judgments and of assessment policies and procedures. A subgroup resolved them through action research. The authors conclude that each of the validity and reliability issues, also identified by the subgroup of multidisciplinary faculty and educational researchers, was also raised by faculty members as they were being trained as assessors. These faculty assessors were from across the disciplines and professions. Thus, faculties experienced in performance assessments who also serve as assessors of broad learning outcomes are likely to continue to develop assessment techniques with appropriate considerations of validity, reliability, and especially consequential validity. At this college, contextual and consequential validity for demonstration of individual student learning outcomes on assessments of integration and transfer imply achievement of complex, multidimensional learning outcomes, so students who were unsuccessful had further opportunity for instruction and reassessment.

Takeaways

- Successful students may be better prepared to succeed in undergraduate professions when they demonstrate integration of their knowledge systems and competences learned and assessed in math and science courses.
- Students are able to adapt and transfer what they have learned on demand in assessments that require them to demonstrate new uses with unfamiliar problems, and that require analysis, problem solving, quantitative literacy, and scientific reasoning.

- Successful students may be better prepared to succeed in undergraduate professions when they demonstrate integration of their knowledge systems and competences learned and assessed in math and science courses.
- Students are able to adapt and transfer what they have learned on demand in assessments that require them to demonstrate new uses with unfamiliar problems, and that require analysis, problem solving, quantitative literacy, and scientific reasoning.

8.1 Introduction

Educational researchers have learned that competent professionals use an extensive knowledge base when engaging their metacognitive processes and using them to create and adapt their performances across various situations they encounter in carrying out their roles (Ericsson et al. 2006; Kane 1992; Mentkowski and Associates 2000; van der Vleuten 1996). However, there seems to be little consensus across disciplines and professions about the nature of relationships between knowledge bases and professional role performances.

Educators engaged in assessment design and establishing validity increasingly experience outside pressures across the educational spectrum. No educators seem to be immune. Secondary schools are critiqued for failing to appropriately prepare students for the workplace (Hout, Elliott, & Committee on Incentives and Test-Based Accountability in Public Education 2011). Undergraduate colleges and universities are under similar pressures to demonstrate critical thinking, problem solving, and communication competencies (Shavelson 2010). Graduate and professional school faculty members often wonder why students who enter are not fully prepared. Graduate professional schools are likewise challenged by societal expectations to prepare graduates who perform their roles competently—and immediately–often on the day they begin their first positions.

Currently, another issue is of major significance. Many of today's professional problems require interprofessional solutions. Often, only experienced professionals meet such expectations. Thus, educational researchers are reviewing opportunities to refine and assess graduates' abilities to define and solve problems they will encounter especially in interprofessional contexts. It is our belief that a greater clarification and consensus across disciplines and professions may lead to more effective professions education overall.

In "Assessing Student Learning Outcomes across a Curriculum", Mentkowski, et al. present a performance assessment instrument to inform both students and faculty about integration and transfer of content and competence across courses and over time. The primary purpose here is to articulate faculty-identified issues that potentially affect the validity and interpretation of results across several learning outcomes, and to show how these issues were resolved.

As authors we examine issues related to defining constructs and resolving issues that emerge when developing assessment and measurement tools to inform teaching/learning, and to generate insights from within and across disciplines and professions. Here are some of the questions:

1. How does a competence-based education model impact learning and assessment?
2. What are some challenges in assessing student learning outcomes across disciplines and professions?
3. What challenges do we face when building consensus definitions of professional constructs for developing and implementing assessments?

8.2 Postsecondary Perspectives and Theoretical Frameworks

Student learning outcomes measures for program evaluation such as the Collegiate Learning Assessment (CLA) are being used more frequently in higher education (see Shavelson 2010). However, faculty at some colleges and universities continues to question whether, for example, the Collegiate Learning Assessment elicits students' best work unless they receive course credit for taking the measure or ensured feedback other than their overall score. Further, faculty members argue that these broad competencies are often not related to a particular institution's curriculum.

While administrators do find evidence from test scores useful for establishing accountability, that same information is less likely to be used by faculty members for curricular improvement because the faculty may find the competencies too broad for identifying specific disciplinary or professional problems. Yet, faculty may agree that the overall findings may be useful for calling attention to broad problems in postsecondary education (Arum and Roksa 2010). However, faculty may also argue that such efforts paint all colleges and universities with the same brush at a time when educators are seeking to identify successful institutions they see as valid comparisons. Further, faculties argue, they are often searching for effective programs that optimize student learning when they look for benchmarks.

In an effort to make program evaluation findings more useful to faculty, the American Association of Colleges and Universities (AAC&U) obtained funding from the U. S. Department of Education to undertake a new project at the same time as the CLA was being introduced in the NASULGC voluntary system of accountability (VSA). The primary goal of the Valid Assessment of Learning in Undergraduate Education (VALUE) project was to elicit performances directly from the curriculum, in particular, using student-created portfolios (AAC&U 2011; Rhodes 2010). The rationale is that the VALUE Rubrics may be applied to complex, multidimensional performances often visible across courses and over time. Portfolios may be the more useful source of valid information for a particular institution's program evaluation efforts.[1]

Thus, to use the rubrics for program evaluation, faculty from across the curriculum may make relationships between what is taught, learned, and assessed in their own curriculum. Faculty then may select VALUE Rubrics appropriate for their own curriculum problems, judge which VALUE Rubrics to use for comparison (goodness of fit), and then apply selected rubrics, dimensions or criteria, and levels to complex student performances. Faculty may also make judgments about which performances demonstrate which dimensions or criteria and levels. Once this evidence is recorded and summarized, faculty can discuss potential improvements to curriculum as well as the consequential validity of the rubrics for their own students (Messick 1989, 1992) (see also Messick 1980, 1982, 1993, 1994, 1999).

In contrast, as authors, we have learned that faculty-designed instruments are more likely to meet the following criteria, even though they are less likely to be used for cross-college comparisons. Those educators involved in cross-college comparisons are aware that this is a difficult and often challenging task (Kuh et al. 2005; Winter et al. 1981; see also Pascarella and Blaich (2013) http://www.liberalarts.wabash.edu/). (Alverno College is also a participant in this study.)

Alverno's gradually evolving criteria for assessment design include:

- The assessment is designed to investigate and resolve questions that are raised by faculty about the quality and effectiveness of student learning. Thus, the rationale for an assessment is problem-based and focused on the learning of individuals and groups.

[1] Rubric authors identify some other purposes as well, for example, using the Rubrics to shape learning outcomes for a series of courses.

- Assessments are specifically connected to courses taught in the faculty's current curriculum because when they are connected, faculty are more likely to make curricular changes that benefit student learning.
- Individual students are not only demonstrating what they have learned in an assessment, they are also learning *while they are completing an assessment*. Challenging tasks across modes are likely to challenge student learning.
- Assessors provide students with individual feedback. Students are challenged and motivated by feedback to do their best work on an assessment.
- Assessors assist students to design their plans for further learning.
- Faculty are motivated because they are designing an assessment to judge whether students can integrate content and competence, that is, knowledge systems and their ability to use them via demonstrated abilities/competencies. Faculty has assurance that students are able to adapt and transfer deep learning to unfamiliar problems in new settings.
- Students complete such instruments in an assessment center, located outside of their classes. The purpose is to create distance from immediate coursework, and in some instances, for students to experience engaging outside assessors from the local professional and business community.
- Faculty members from across the disciplines and professions are trained as assessors and are interested in the results and in ensuring all students succeed.
- Educational researchers are motivated because they are working side by side with faculty in using disciplinary and professional learning principles (for example: *learning is integrative and transferable)* that are connected to assessment principles (for example, *assessment requires integration of knowledge systems and competencies, and their transfer across assessment modes and contexts unfamiliar to students)* (Bransford et al. 2000; Mentkowski et al. 2000).

8.3 Purposes and Problems to Consider for Developing a New General Education Assessment Across Disciplines and Professions

One purpose of this chapter is to examine the assumption that investments and benefits of performance assessments for undergraduates develop as campus faculty—across disciplines and professions—design instruments for assessment center administration outside-of-classes. It is completed by undergraduates after they complete general education in the liberal arts and before they enter the professions. The Alverno schools of nursing, education, business and management, and arts and technology prepare students in such diverse professions as nursing, education, business and management, and professional communications.

The broad purpose of this faculty-designed assessment technique described here is to assess for integration and transfer of student learning outcomes across selected prior coursework in general education and over time. In this particular case, the assessment is for judging levels of scientific reasoning, analysis, problem solving, and quantitative literacy. Students have already successfully demonstrated these learning outcomes on in-class assessments in math and science courses.

8.3.1 Nature of the Faculty-Identified Problem

Several observations precipitated action on the part of the Council for Student Assessment and its External Assessment Subgroup (EAS). (This group is responsible for studying and validating external assessments conducted by the Assessment Center.) For some time, the EAS had been working on the problem of students' experiences with external assessments, noting from a student survey that some students questioned the value of an external assessment when they had already demonstrated "the same things" in classes. The EAS was also challenged by their colleagues in the Self Assessment Subgroup, who wondered if some changes might be undertaken to strengthen self assessment for students by asking them to self assess across their prior coursework, using assessor feedback. A larger effort was underway to engage students in more rigorous and demanding experiences, following completion of general education, that is, to provide an assessment with more challenging content and abilities. In this context, the EAS began to grapple with how integrated learning of Alverno Abilities transferred to role performance for current students. Taking this action was in contrast to blaming the students who had learned in these earlier courses and the faculty who had taught and assessed their performances.

Thus, the EAS recruited five math and science faculties to examine the problem of better assisting students to integrate their course content and abilities, adapting them, and then transferring their knowledge and abilities to unfamiliar settings in their major field or profession. The faculty design team relied on findings from several sources. First, the design team relied on their own experiences, teaching, and assessing in their classes with current students. Since faculty were experienced in developing performance assessments for their courses, they were curious about whether the problem of integration and transfer could be resolved in part by examining student performance following completion of general education coursework. Their own experiences were validated when Alverno educational researchers found that some students recycle to earlier forms of understanding what they had earlier been able to demonstrate—such as ways of thinking about their knowledge and abilities—when they were faced with unfamiliar problems in unfamiliar contexts (Mentkowski 1988).

Alverno researchers had also shown that most five-year alumnae were capable of such adaptation and transfer of integrated content and abilities (Mentkowski and Associates 2000). Experienced faculty members on campus were aware of these findings. Third, some faculty on the EAS were aware of findings in the broader literature that directly named the problem of lack of integrative learning and transfer (Bransford et al. 2000; Huber and Hutchings 2004; Mentkowski and Sharkey 2011).

Perhaps the most influential issue for the EAS and the most persuasive for the design team was that faculty members who were teaching courses in the professions observed that students too often did not demonstrate what they had learned in earlier math and science courses when they entered courses in their major field, which for many students, were their courses in the professions. This is often a common problem in higher education. They communicated this to their colleagues in various ways: cross-department meetings, ability department meetings, and all-faculty meetings. One faculty member went so far as to say in a public meeting that: "Too many of our students avoid using quantitative evidence to make arguments, even when it is right in front of them."

Consequently, three faculty members, Johanson, Mente, and Young, who were teaching STEM disciplines, in this case science and mathematics, and two experts in assessment (Abromeit, Chair of the Council for Student Assessment, and O'Brien-Hokanson, former Co-Chair) created an assessment design. Together, team members challenged each other to set a higher bar for student learning: Optimize the likelihood that students are able to use prior coursework in subsequent classes in the major fields and professions.

Thus, the assessment design would assess for individual student learning and demonstration of *integrating content with competence across courses and over time.* In this case, the assessment design would examine whether and how individual students integrate, adapt, and transfer scientific reasoning and quantitative literacy as they analyzed and solved unfamiliar problems in a setting they had not experienced before. Thus, their integrated abilities of scientific reasoning, quantitative literacy, analysis, and problem solving were to be demonstrated beyond when and where students had learned to integrate these concepts and competencies—in their courses, where they were also taught and assessed. The Alverno curriculum requires all students to demonstrate eight abilities (competence integrated with content) in order to graduate. Faculty members hold students to criteria in their courses. Criteria are criterion-referenced and often norm-referenced.

A first task for the design team was to identify learning outcomes for the assessment that were intentionally connected to students' prior coursework, to develop the strategies necessary to draw their colleagues' attention to these courses, and to gradually put in place a series of course requirements that students would complete prior to completing the assessment. In the past, EAS had noticed that some students procrastinated on taking their external assessment in general education, and a few finally completed it during their senior year. EAS determined that students who did so were not availing themselves of the opportunity to demonstrate integration and transfer, nor would they receive the extensive feedback on their performance or assistance with planning for their further learning.

8.4 Methods

Description of Alverno assessment process as assessment-as-learning is multifaceted and *connects directly to student learning.*

- When learning is developmental, it results in new thinking/understanding (Alverno College Faculty 1985, revised 1994, p. 17).
- The learners' developing capacity to integrate knowing and doing in this curriculum with reflective awareness is transformative (Mentkowski and Associates 2000, p. 237).
- An instrument/prompt designed to assess self assessment must "look forward beyond the level at hand," i.e., "elicit the most advanced performance of which she is capable." (Alverno College Faculty, p. 23).
- "Prior levels of ability must not only be reinforced but also be drawn into more and more complex uses" (Alverno College Faculty, p. 23).
- "At least some of the criteria used in appraisal are *fuzzy* rather than *sharp*…A *fuzzy* criterion is an abstract mental construct denoted by a linguistic term which has no absolute and unambiguous meaning independent of its context. If a student is able to consciously use a *fuzzy* criterion in making a judgment it is necessary for the student to understand what the *fuzzy* criterion means and what it implies for practice. Therefore, learning from these contextualized meanings and implications is itself an important task for the student (Sadler p. 119)."

Assessor training for faculty assessors and establishing validity of training fifty-two faculty members from across the disciplines and professions (humanities, natural sciences and mathematics, social and behavioral sciences, arts and technology, nursing, business and management, education, professional communications, community leadership) plus seven assessment center staff participated in training conducted by EAS team members (natural scientist and social scientist). Assessor training sessions occurred over an academic year. Action research involved assessors in processes designed to improve assessor training as a result of their instruction.

A member of the faculty with both the degree in the humanities and in the social sciences commented:

> The assessor training was straightforward. I left the training with an understanding of the overall purpose of the assessment, a sense of how the parts likely flowed together, and exposure to the assessment materials. I felt fairly confident.
>
> I enjoyed assessing [Mid-program General Education Assessment] AC 309. The assessment flowed logically and materials were easily accessible. The students' reflection on their prior work was interesting. It was fun witnessing our students applying their quantitative reasoning skills in such a practical manner. Although I was one of the last to finish this first time through, I already feel confident that I will not continue to be among the last to complete my feedback! I look forward to assessing AC 309 again.
>
> Brenda Kilpatrick, MA (Theological Studies), MA (Clinical Psychology), Assistant Professor, Department of Psychology

8.4.1 Data Sources for Establishing Design-Based Validity and Reliability

An independent evaluator recorded faculty questions during training sessions. Mentkowski used a combination of deliberative inquiry (Harris 1991), qualitative methods, and action research (O'Brien 2001; Reason and McArdle 2008) to examine assessor training procedures, the construct validity of the instrument, and the procedures for instrument administration. For example, during each of the # training sessions she observed, Mentkowski identified faculty questions related to (1) the purposes of general education assessments for student learning; (2) student learning outcomes that were being assessed with the technique; (3) procedures for the administration of the assessment technique by the assessment center; (4) the faculty assessor role during the assessment; (5) purposes for particular procedures during the assessment process; and (6) Alverno assessment policies related to how students were learning *during the time they were completing the assessment*.

Most Alverno faculty members had been previously trained as assessors. Each assessor was recruited by his or her dean or associate dean from across the disciplines and professions. They willingly signed up for training. Thus, faculty questions *did not occur* in the categories named above. Rather, questions from individual assessors emerged during discussions, and further questions emerged as the assessor training progressed over several sessions. Consequently, Mentkowski recorded the responses by trainers Abromeit and Mernitz as they were conducting the training.

Mentkowski noticed that questions were often raised about policies and procedures during the assessor training sessions. Often, the trainers responded to these questions. In some instances, trainers suggested they would also raise the questions at the next meeting of the EAS. Using qualitative research procedures, Mentkowski then translated these faculty assessor questions into declarative sentences and statements. She noticed that when she formulated these questions as statements and articulated questions, trainer answers, and other statements by faculty assessors, these declarative statements would elicit debate and discussion at the weekly team meetings of the EAS, held prior to subsequent assessor training so that training could be improved. This particular method qualifies as action research. The EAS consistently made revisions to improve the construct validity of the instrument (based on faculty questions raised across the disciplines and professions), and the procedures for instrument administration (fairness issues for both students and faculty assessors).

The independent evaluator, Mentkowski, also communicated reliability findings, as these became available, during assessor training sessions. Mentkowski did not serve as a trainer of assessors, a coach, or an assessor.

8.4.2 Evidence for Faculty-Identified Validity Issues

During the assessor training sessions, faculty began to engage each other interactively—across the disciplines and professions—about what it meant to resolve validity issues regarding their all-too-human judgments related to criteria (Hammond 1996; Mentkowski and Rogers 1988; Rogers 1994) (see also Mentkowski and Rogers 1985, 1986). Not only did faculty members begin to insist on studying the consistency and reliability of their judgments (with a little help from their colleagues in the psychology department and the School of Nursing). Faculty also elaborated on the basis for judgments, assisted by faculty in the STEM disciplines. However, other disciplines also weighed in on the basis for judgment that led to several clarifications in what it meant for students to provide a null hypothesis (with the assistance of the psychology department), and what it meant for students to provide a declarative sentence as a hypothesis statement (led by the Biology Department). The resolution to this issue was cross-disciplinary. Assessors were instructed during subsequent training sessions that both types of hypotheses were to be acknowledged as "*met criteria*."

8.4.3 Evidence for Reliability of Faculty Judgment

Assessors explained the judgments they made about the student's performance, to the student, indicating where the assessor decided the performance *met criteria, partially met criteria,* or *did not meet criteria.* The assessor role was not to begin to instruct students in the integrated constructs and competencies/capabilities, but rather to communicate to the student how they made an overall judgment of *succeed* or *did not succeed* based on the pattern of performance the student demonstrated: *met, partially met, did not meet.* Whether or not the student's performance succeeded, the assessor engaged the learner in creating a plan for further learning. Thus, students whose performances were unsuccessful were assisted to use assessor feedback to develop a plan for further learning. Thus, they received a benefit that would be engaging to the student and communicate to him/her that he/she would be continuing to learn. The authors do not claim that this benefit is similar to that of the successful student, however, both successful and unsuccessful student performances are communicated to students. Each receives a similar message from the assessor: learning is a continuous process.

8.4.4 Establishing Validity of Assessor Final Judgment and Fairness to Each Student

When a student did not succeed, an experienced coach present during the assessment provided an independent view of the performance, usually by taking the assessor through his or her decision-making process for arriving at a final judgment. Thus, a student knew when she left the assessment whether she had succeeded or not. This is evidence for procedural fairness to students.

8.4.5 Evidence for the Statistical Reliability of Faculty Judgment

Inter-judge agreement was 100 % for "did not succeed" because an experienced coach was available to verify an assessor's individual judgment. Further, independent-judge agreement for a random sample (n = 40) of summative judgments was 95 % for "succeeded and did not succeed." Methods also included systematic data collection of 204 performances, to establish performance-based validity.

8.4.6 Evidence for Consequential Validity of Assessment Policies and Procedures

Faculty assessors during the training sessions began to discuss issues related to consequential validity as defined by Messick (1994). As a result of discussions about the consequences of the assessment for individual students, in particular by the humanities faculty assessors, the EAS decided that students who did not meet criteria would be invited to an intervention workshop.

As noted earlier, all students who do or do not succeed on the assessment are requested to develop a picture of strengths and weaknesses and a plan for further learning. The Subgroup argued that students who had developed their plans would be invited to a follow-up workshop, so that they would achieve the benefits of further instruction, and also be assessed. So far, all of the students who did not succeed attended the intervention workshop. Afterward, students completed a shorter version of an assessment with different problems. So far, only two students have not passed the next assessment.

8.4.7 Feedback from Assessors

Following completion of the assessment by students, faculty assessors met with each student to provide individual feedback. Faculty strove to provide *accurate, conceptual, diagnostic, inferred from performance, and prescriptive* feedback.

8.4.8 Intervention Workshops

If a student fails to demonstrate criteria on the general education assessment, she is required to participate in an intervention workshop prior to attempting a reassessment. The goals of the workshop are to strengthen students' preparation strategies for the general education assessment, to strengthen students' abilities related to quantitative literacy levels 1 and 2, to strengthen students' abilities related to scientific reasoning, quantitative literacy integrating analysis, and problem solving, and to strengthen students' understanding of the connection between performance and self assessment. During the intervention workshop, students are also given an opportunity to share insights related to their experience with the assessment process and the assessment itself.

Students complete a reflection on their preparation for the general education assessment and on the ability areas that caused them the most difficulty on the assessment prior to attending the workshop. This serves as a basis for group discussion and provides direction for the workshop. At the beginning of the workshop session, peers share insights, effective preparation strategies, and often strategies they wish they had employed; facilitators provide advice for preparing for reassessment.

The majority of the workshop consists of students completing portions of scientific reasoning and quantitative literacy-based activities using a combination of instructional approaches. The activity components that are completed during the workshop are determined by the beginning discussion of what each student found difficult on the original assessment. Facilitators lead discussions with some reteaching of concepts as needed. Students practice ideas in peer groups, and groups report struggles and strategies to add support for each other and to learn from each other. The session is dynamic and interactive.

At the end of the workshop, all students are provided with a solution set for the entire activity, including questions that were not used during the workshop session. They are also provided with another scientific reasoning and quantitative literacy activity, with analytical problem solutions to use for practice on their own to reinforce the strategies developed during the workshop.

A representative from the assessment center was available at the end of the workshop session to schedule reassessment opportunities for the students. These are scheduled individually for each student. All students must successfully complete a reassessment to meet the general education requirement for graduation.

8.5 Summary of Findings

Validity issues included: (1) achieving clarity of purpose for out-of-class assessments (integration of knowledge and competencies and their adaptation and transfer); (2) relationship between knowledge/skills assessed and courses completed; (3) who completes the assessment (two-year students and students who entered from other colleges); (4) whether policies and procedures were rigorously reasoned and fair to students and assessors; (5) whether summative assessor judgments were reliable; (6) whether consequences were appropriate for both students who succeeded and those who did not succeed, given faculty learning and assessment principles; (7) who infers recommendations from analzses of student performances and implications for curricular improvement (assessment council or general faculty); and (8) whether program evaluation and student learning purposes compete (Mentkowski 1991, 1998, 2006; Loacker and Rogers 2004). Because not all students succeeded (77 %) the instrument can be used for program evaluation of integration and transfer of content with competence, but not, argued faculty, without intervention workshops and comparable reassessments for ensuring mastery.

8.5.1 Noticing Intentional Changes in the Curriculum

During the faculty assessor training, faculty members across the disciplines began to look for ways to improve teaching and learning at the department level. (1) They raised questions about where STEM departments taught for the integration of content and competence, and STEM faculty members articulated where students engaged in coursework to meet the general education assessment expectations. As a result, some STEM faculty began a discussion with non-STEM faculty whether such competences, especially quantitative literacy, were taught and reinforced in other courses other than STEM.

As a result of the general education assessment, other faculty across the disciplines and professions began to take responsibility for teaching scientific reasoning and quantitative literacy as they were also developing students' analysis and problem solving competencies.

Staff Assessor Jill Haberman took the following notes at the Communication Ability Department meeting on January 18, 2012, where faculty and academic staff raised issues about the general education assessment following the November 18, 2011 presentation of the results to the entire faculty.

- Dr. Kevin Casey (Professor of History) commented that a number of conversations have been going on in the Humanities Division as instructors are experiencing assessor training that challenges their own comfort levels in quantitative literacy. They are questioning how they might better prepare students to analyze and present statistical information since students are now being called to do so in the general education external assessment. Casey specifically mentioned graphing, and also said that the History department already has an agenda of ideas on this. Casey asked the Communication Ability Department Subcommittee on Quantitative Literacy to reach out to the Humanities faculty to help with these ideas and Suzanne Mente (Assistant Director of Instructional Services) agreed to do so.

- Mente commented that those course instructors who teach and assess for Quantitative Literacy have been made aware of the struggles that some students have had in demonstrating Quantitative Literacy levels 1 and 2. Currently, these instructors are using the specific findings from the general education assessment [presented to the entire faculty on November 18, 2011] to adjust their instruction, in particular, assisting their students in describing and comparing using quantitative data.
- Dr. Robert O'Brien-Hokanson (Professor of English) mentioned that Integrated Communication Seminar instructors have been discussing ways to help students to better describe patterns [in quantitative evidence].
- Dr. Susan Pustejovsky (Professor of Mathematics) noted that the general education external assessment presents serious reading challenges for some students, particularly the charts.
- Dr. Robert O'Brien-Hokanson also commented that he plans to address "Reinforcement of Quantitative Literacy Skills" in a meeting with Integrated Communication Seminar instructors.[2]

8.6 Conclusions and Implications for Professions Education

Assessors in a wide range of disciplines may prioritize values differently (autonomy; open-mindedness) than professions (teamwork, client service). Yet faculty in this integrated liberal arts and professions curriculum reached consensus on assessment purposes for a general education assessment. Humanities faculty challenged consequential validity (Do majors in humanities need scientific reasoning? What policies support unsuccessful students?), and agreed to be assessors because faculty themselves should demonstrate general education outcomes for students. When humanities learning outcomes are integrated into professions curricula, assessments should demonstrate consequential validity.

8.7 Scientific and Scholarly Value

Successful students may be better prepared to succeed in undergraduate professions when they demonstrate integration of their knowledge systems and competencies learned and assessed in math and science courses. However, whether students are

[2] Statements were checked by participants from the meeting to ensure accuracy.

able to adapt and transfer what they have learned on demand in assessments that require them to demonstrate new uses with unfamiliar problems that require analysis and problem solving is a question for faculty across higher education.

The Collegiate Learning Assessment and the VALUE Rubrics created by AAC&U are strategies that work for program evaluation. However, ultimately, faculty-designed assessments that are (1) directly related to prior coursework in a curriculum, (2) administered outside-of-class with unfamiliar problems that require students to integrate knowledge systems and competencies, and adapt and transfer their learning to a new setting, we believe, are important to ensure that undergraduates in the liberal arts are ready to enter the undergraduate professions.

Issues/Questions for Reflection

- How does a competence-based education model impact learning and assessment?
- What are some challenges in assessing student learning outcomes across disciplines and professions?
- What challenges do we face when building consensus definitions of professional constructs for developing and implementing assessments?
- How does a competence-based education model impact learning and assessment?

References

Alverno College Faculty. (1985, revised 1994). *Student assessment-as-learning at Alverno College.* Milwaukee, WI: Alverno College Institute.

American Association of Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' views*. Washington, DC: American Association of Colleges and Universities.

Arum, R., & Roska, J. (2010). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press: National Research Council.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, and unavoidable injustice*. New York: Oxford University Press.

Harris, I. B. (1991). Deliberative inquiry: The arts of planning. In E. C. Short (Ed.), *Forms of curriculum inquiry* (pp. 285–308). Albany: State University of New York Press.

Hout, M., Elliott, S. W. (Eds.). (2011). Committee on Incentives and Test-Based Accountability in Public Education. *Incentives and test-based accountability in education.* Washington, DC: National Research Council.

Huber, M. T., & Hutchings, P. (2004). *Integrative learning: Mapping the terrain*. Washington, D. C.: American Association of Colleges and Universities.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*(2), 163–182.

Kuh, G. D., Kinzie, J., Schuh, J. H., Whitt, E. J., et al. (2005). *Student success in college: Creating conditions that matter*. San Francisco: Jossey-Bass.

Loacker, G., & Rogers, G. (2005). *Assessment at Alverno College: Student, program, institutional*. Milwaukee, WI: Alverno Institute.

Mentkowski, M., & Rogers, G. P. (1985). *Longitudinal assessment of critical thinking in college: What measures assess curricular impact?*. Milwaukee, WI: Alverno College Productions.

Mentkowski, M., & Rogers, G. (1986). Assessing critical thinking. In L. S. Cromwell (Ed.), *Teaching critical thinking in the arts and humanities* (pp. 117–128). Milwaukee, WI: Alverno Productions.

Mentkowski, M., & Rogers, G. (1988). *Establishing the validity of measures of college student outcomes*. Milwaukee, WI: Alverno College Institute.

Mentkowski, M. (1988). Paths to integrity: Educating for personal growth and professional performance. In S. Srivastva, et al. (Eds.), *Executive integrity: The search for high human values in organizational life* (pp. 89–121). San Francisco: Jossey-Bass.

Mentkowski, M. (1991). Creating a context where institutional assessment yields educational improvement. *Journal of General Education, 40*, 255–283. (Reprinted in *Assessment and program evaluation* (ASHE Reader Series), pp. 251–268, by J. S. Stark, & A. Thomas, Eds., 1994, Needham Heights, MA: Simon & Schuster).

Mentkowski, M. (1998). Higher education assessment and national goals for education: Issues, assumptions, and principles. In N. M. Lambert & B. L. McCombs (Eds.), *How students learn: Reforming schools through learner-centered education* (pp. 269–310). Washington, DC: American Psychological Association.

Mentkowski, M., & Associates (2000). *Learning that lasts: Integrating learning, development, and performance in college and beyond*. San Francisco: Jossey-Bass.

Mentkowski, M. (2006). Accessible and adaptable elements of Alverno student assessment-as-learning: Strategies and challenges for peer review. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 48–63). London, UK: Taylor and Francis.

Mentkowski, M., & Sharkey, S. (2011). How we know it when we see it: Conceptualizing and applying integrative and applied learning-in-use. In Jeremy D. Penn (Ed.), *Assessing complex general education student learning outcomes, (*pp. 48—63). *New Directions for Institutional Research* (149, Spring). San Francisco: Jossey-Bass.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.

Messick, S. (1982). *Abilities and knowledge in educational achievement testing: The assessment of dynamic cognitive structures*. Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments*. Princeton, NJ: Educational Testing Service.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61–73). Hillsdale, NJ: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. J. (Ed.). (1999). *Assessment in higher education: Issues of access, quality, student development, and public policy*. Mahwah, NJ: Erlbaum.

O'Brien, R. (2001). Um exame da abordagem metodológica da pesquisa ação [An overview of the methodological approach of action research]. In Roberto Richardson (Ed.), *Teoria e Prática da Pesquisa Ação [Theory and practice of action research]*. João Pessoa, Brazil: Universidade Federal da Paraíba. English version retrieved 5/25/11 at http://www.web.ca/~robrien/papers/arfinal.html

Pascarella, E. T., & Blaich, C. (2013). Lessons from the Wabash National Study of Liberal Arts Educaton. *Change: The Magazine of Higher Learning*. London: Taylor and Francis.

Reason, P., & McArdle, K. L. (2008). Action research and organization development. In T. Cummings (Ed.), *Handbook of organization development* (pp. 123–137). Thousand Oaks, CA: Sage Publications. Retrieved 5/25/11 at http://www.peterreason.eu/Papers/ActionResearch&OrganizationDevelopment.pdf

Rhodes, T. L. (Ed.). (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: American Association of Colleges and Universities.

Rogers, G. (1994, January–February). Measurement and judgment in curriculum assessment systems. *Assessment Update*, 6(1), 6–7.

Rogers, G., & Mentkowski, M. (1994). *Alverno faculty validation of abilities scored in five-year alumna performance*. Milwaukee, WI: Alverno College Institute.

Rogers, G., & Mentkowski, M. (2004). Abilities that distinguish the effectiveness of five-year alumna performance across work, family, and civic roles: A higher education validation. *Higher Education Research & Development, 23*(3), 347–374.

Sadler, R. D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41–67.

Winter, D. G., McClelland, D. C., & Stewart, A. J. (1981). *A new case for the liberal arts: Assessing institutional goals and student development*. San Francisco: Jossey-Bass.

Chapter 9
Assessing Across Domains of Study: Issues in Interdisciplinary Assessment

Mark Russell, Anne McKee and Michele Russell-Westhead

Abstract In the United Kingdom, the design and process of university assessment, particularly at the undergraduate level, has come under intense public scrutiny following the implementation of policy aimed at improving educational standards. The introduction of student fees and national satisfaction surveys has promoted competition between institutions for student numbers with 'assessment and feedback' being considered a key indicator of student engagement and overall satisfaction with their experience. This has put assessment at the centre of learning and teaching strategies and innovation. Performance assessment in professional education, particularly in medicine and nursing, is primarily focused on ensuring fitness to practice and has more well defined performance outcomes than more traditional science and social science disciplines. Healthcare students are required to keep a portfolio of knowledge and skills supported with competence assessments. Other disciplines in higher education are now required to keep student portfolios of transferable skills as part of demonstrating 'employability'. Undertaking formative feedback on student performance, particularly in the work place, is becoming more challenging. Increasingly in the multi-disciplinary and high-service demand environments of the National Health Service, clinical care priorities leave little time for educational activities. In response, profession educators are re-conceptualising forms of feedback and assessment in these contexts. Examine ways in which assessment strategies and methods need to evolve in clinical and higher education to respond to policy, professional body and student expectations, and the implications of this for faculty development in both institutional and work-based contexts. Drawing upon national and institutional data, we have identified key assessment issues, notably the importance and challenge of receiving timely and purposeful feedback. We also offer suggestions for strategic and innovative solutions, for example, advantages of creating curriculum as opportunities for student learning, but viewing curriculum as a set of system-wide development processes.

M. Russell (✉)
King's Learning Institute, King's College London, London, UK
e-mail: mark.russell@kcl.ac.uk

A. McKee · M. Russell-Westhead
King's College London, London, UK

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_9

Takeaways

- Governmental policies relating to higher education can shape how universities articulate their educational missions and inform their educational governance and practices. Such policy initiatives can also strengthen the importance of education, even in research intensive universities.
- A case study within one university in the United Kingdom (U.K.) to improve assessment across disciplines and professions highlighted the importance of institutional and local leadership to enable change across the organization. Local leaders need to have sufficient time to direct, plan, evaluate and adapt their assessment improvements within their academic departments.
- This distributed form of institutional leadership takes account differing priorities of the participating groups, allowing change to be owned by faculty, clinical teachers, students and managers.
- There is much to be gained by stakeholder engagement, within and across the participating groups. Such stakeholders include faculty, students, professional bodies, quality assurance teams and, in health-care settings, clinical teachers.
- Instruments/tools/techniques and processes should be developed to help participating groups inquire into their own practice and not be presented with solutions they have little ownership of. Typically top-down approaches fail to understand and respond to local contexts.

9.1 The Higher Education Policy Context

In the United Kingdom (UK), the design and process of university assessment, particularly at the undergraduate level, has come under intense public scrutiny following the implementation of policy aimed at improving educational standards. A series of Higher Education Funding Council (HEFCE) national initiatives to improve the quality and status of learning and teaching simply disappointed the sponsor. The speed and scale of change was slow and patchy. Encouraging the sector to move in particular directions proved a more complex negotiation than the voluntary engagement of teaching enthusiasts. What was happening? Why was the task of improving learning and teaching too complex? Were the resources insufficient or ineffectively deployed? Perhaps all of these. In addition, there was an ideological tension that was to influence how higher education was to evolve.

Underpinning concerns about the quality of higher education lay fundamental shifts in thinking about the role and purpose of higher education. These involve contested views about what learning, teaching and assessment could and should involve. The international and national dimensions of higher education policy provide a significant context for understanding the values, ideologies and aims within the assessment debate.

In the late 1990s and early twenty-first century, there was a global focus on the need for a knowledge-based society (Garnham 2002). Fears about economies failing were central to debates about the educational implications for developing learners to meet the needs of a knowledge-based society. A prevailing view argued that technological innovation, the improved speed and access of communications, and the growth of global markets required a new kind of workforce. That workforce was characterised by workers who could quickly re-skill, acquire new knowledge and move into changing roles.

Preparing students to meet the challenge of becoming part of such a workforce requires an ability to adapt to rapid and frequent change. This is not a new challenge for education in the professions tasked with preparing practitioners for the complex and dynamic environments of health-care settings. The 2013 the final report of an independent review of postgraduate medical training called '*The Future Shape of Training*' identifies similar workforce needs with a particular emphasis on flexibility to undertake new roles.

In professional and higher education, the intensified focus on workforce planning, and employability has pedagogic implications. Dr. Nick Hammond, senior adviser at the Higher Education Academy in the UK, describes these below:

> The changing world to be faced by today's students will demand unprecedented skills of intellectual flexibility, analysis and enquiry. Teaching students to be enquiring or research based in their approach is central to the hard-nosed skills required of the future graduate workforce (Hammond 2007, p. 3).

There was more at stake than shifting pedagogic approaches. The purpose and functioning of universities, the social contract between higher education and the state, was being re-drawn. Responsibilities of universities to contribute to the well-being of society were being reconsidered. Universities were rethinking access. This was affecting what universities did, for whom, and how. This had serious implications for individual and institutional identities (McKee 2012).

UK institutions have created employability opportunities and associated support mechanisms for their students. Some have gone further and oriented themselves firmly toward an employability agenda. Just as other institutions use their research pedigree to distinguish themselves, some universities are using their employability focus to create an institutional identity that distinguishes themselves from their competitors. [For example: Coventry University (www.coventry.ac.uk), the University of Hertfordshire (www.herts.ac.uk), and The University of Plymouth (www.plymouth.ac.uk).]

UK policy changes affected both universities and students. The funding of universities and their degrees was at stake. The government paid student fees directly to universities and provided students with grants to cover living costs while they studied. This policy was transformed in stages. In 1990, The Education

(Student Loans) Act was passed in the UK which meant students now received and repaid loans towards their maintenance while studying.

Student loans were part of a series of UK policy initiatives to contain public spending on higher education. The shift from giving students grants to providing student loans occurred in 1997 with the *Dearing Report.* Since then, the cost of university study in England has trebled from an average of £3000 per student per year to a capped figure of £9000 per student per year. This increase in fees created two concerns within the sector. The first was that access, especially for low-income families, would be compromised. The second was that student expectations would rise.

Parallel with changes in resourcing levels and funding structures were a series of UK Government initiatives to improve the quality of learning and teaching. These involved a shift from a *non-intervention* approach that highly valued the autonomy of universities to *categorical funding* to generate development in broad areas, such as *enterprise*. One example was the creation of a *National Teaching Fellowship Scheme*. This award helped celebrate individual teaching excellence. Another initiative, *Funds for the Development of Teaching and Learning*, offered resources to develop good teaching practice in priority areas, such as *inter-professional education.*

Other initiatives sought to enhance the structures supporting teaching. Twenty four '*Subject Centres*' of broadly grouped discipline areas were created to focus work within disciplines. Subject Centres were hosted by institutions, distributed across the UK, with oversight and control by the UK *Higher Education Academy*. The final initiative was a 315 million pound sterling five-year programme to create institutionally based *Centres for Excellence in Teaching and Learning* (CETL). Seventy three CETLs were funded and tasked with having impact at an institution, and across a sector. Some CETLs were focused on disciplines (*Centre for Excellence in Mathematics* and statistics support), some on areas of activity (*Assessment for Learning* CETL), and others on institutional imperatives (e.g. *The Blended Learning Unit*).

However, few of these initiatives have been sustained. Improving the quality of learning, teaching, and assessment is a complex activity requiring more time, resources, and widespread collaboration.

The UK *National Student Survey* (NSS) has been influential in focusing institutional development, particularly in areas of assessment and feedback. Launched in 2005, all undergraduate students complete it in their final year. *NSS* is a retrospective evaluation by students of their experience in eight categories:

- Teaching On My Course,
- Assessment and Feedback,
- Academic Support,
- Organisation and Management,
- Learning Resources,
- Personal Development,
- Overall Satisfaction, and
- Students' Union.

Sponsors of the survey, The *Higher Education Funding Council for England* (HEFCE), intended that information gathered would give students a voice in the quality of their learning and teaching experience. The information is publically shared. Prospective students and their parents and families make choices for particular degree programmes and universities.

The survey's validity, how it respects the complexity of the student experience, and its reliability across time, has been hotly debated. Critics argue that student engagement has been neglected. Developers of the survey want to better understand learner's experience away from the current consumer orientation toward a student engagement approach. Students are not merely passive learners but have a role and responsibility in their own learning journey. For example, the *National Survey of Student Engagement* (NSEE) used in the United States is an engagement rather a satisfaction-oriented survey. Despite its focus, the annual production of league tables of student satisfaction has made the *NSS* both influential and consequential. Across the sector, student satisfaction in the *Assessment and Feedback* category tends to be consistently low. Though some universities have made significant improvements, assessment and feedback is problematic across much of the sector. Though *NSS* might be helpful in identifying an area of concern, it offers scant information and little if any insight into the nature of the challenge.

The *Assessment Standards Knowledge Exchange* (ASKe), one of the Centres for Excellence in Learning and Teaching, convened an international group of academics and assessment experts to probe the complexity of *Assessment and Feedback* to help support change across disciplines. In 2007, a convened group (*Western Manor Group* of 30 participants) produced *Assessment: A Manifesto for Change*. The Manifesto focuses on assessment standards (one of the tenants is *Standards Without Standardisation*) and is comprised of six separate-yet-linked tenets. These are:

1. The debate on standards needs to focus on how high standards of learning can be achieved through assessment. This requires a greater emphasis on assessment for learning rather than assessment of learning.
2. When it comes to the assessment of learning, we need to move beyond systems focused on marks and grades towards the valid assessment of the achievement of intended programme outcomes.
3. Limits to the extent that standards can be articulated explicitly must be recognised since ever more detailed specificity and striving for reliability, all too frequently, diminish the learning experience and threaten its validity. There are important benefits of higher education which are not amenable either to the precise specification of standards or to objective assessment.
4. Assessment standards are socially constructed so there must be a greater emphasis on assessment and feedback processes that actively engage both staff and students in dialogue about standards. It is when learners share an understanding of academic and professional standards in an atmosphere of mutual trust that learning works best.

5. Active engagement with assessment standards needs to be an integral and seamless part of course design and the learning process in order to allow students to develop their own, internalised, conceptions of standards and monitor and supervise their own learning.
6. Assessment is largely dependent upon professional judgement and confidence in such judgement requires the establishment of appropriate forums for the development and sharing of standards within and between disciplinary and professional communities.

In 2009, ASKe convened another forum (*Osney Grange Group*) to identify and examine key issues in feedback and suggest possible resolutions. *Feedback: An Agenda for Change* was a significant outcome from the Group. This Agenda sets forth "the intention of tackling underpinning theoretical and practice issues in feedback and creating a cross-disciplinary framework to inform improvement and development" (see Merry et al. 2013).

The five clauses of *Feedback: An Agenda for Change* are:

1. High-level and complex learning is best developed when feedback is seen as a relational process that takes place over time, is dialogic, and is integral to learning and teaching.
2. Valuable and effective feedback can come from varied sources, but students must learn to evaluate their own work or they will be dependent upon others to do so. Self and peer review are essential graduate attributes.
3. There needs to be a fundamental review of policy and practice to move the focus on feedback from product to process.
4. Reconceptualising the role and purpose of feedback is only possible when stakeholders at all levels in higher education take responsibility for bringing about integrated change.
5. The Agenda for Change calls on all stakeholders to bring about necessary changes in policy and practice (Price et al. 2013).

What are the implications of this policy context for institutions? We address this question by examining a *Feedback and Assessment* evaluated initiative within 'King's College London'.

9.2 King's College London: An UK Higher Education Institution

King's College London is a large UK university located in the heart of London, England. Approximately 25,000 students study at King's and are taught in one of eight academic Faculty's. The University has a significant emphasis on research and is a member of the UK elite Russell Group. As such, any significant endeavours need to be cognisant of the research-intensive context. Although King's is highly ranked in international university league table some student satisfaction league tables highlight

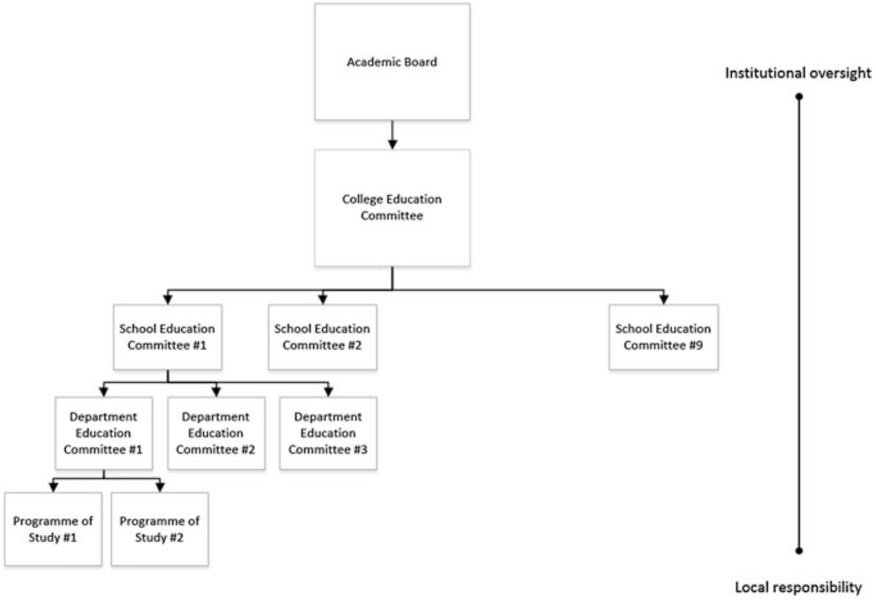

Academic Board
College Education Committee
School Education Committee #1
School Education Committee #2
School Education Committee #9
Department Education Committee #1
Department Education Committee #2
Department Education Committee #3
Programme of Study #1
Programme of Study #2
Institutional oversight
Local responsibility

Fig. 9.1 Overview of educational governance at King's College London

less favourable ranking. Additional information relating to the institutional context and history of King's College London can be found in Appendix.

The governance of education at King's is highlighted in Fig. 9.1. There is a clear connection from academic programmes, departments, to academic Faculties, to an overseeing College Education Committee. Such connections ensure a university connectedness and allow the educational values and strategies to flow throughout the institution. The responsibility for ensuring faculty-based academic standards and driving continual educational improvements is devolved to a dedicated faculty-based education lead. The College Education Committee, of which each of the Faculty education leads are members, is chaired by the University's lead for education [Vice Principal (Education)].

The University has an Education Strategy which sets out the strategic imperatives and presently sets out, amount other things, the importance of technology-enhanced learning, assessment and feedback.

9.3 A University Initiative

Prevailing and often reiterated NSS concerns about assessment and feedback, and our understanding of the importance of assessment for student learning, has led to a cross-University, two year, assessment and feedback project.

9.3.1 Guiding Principles

- Ensure that teaching, learning outcomes, and assessment are constructively aligned (Biggs 2003).
- Create an institutional imperative while allowing scope for local-relevance.
- Engender responsibility for enhancement of assessment, as appropriate, with the institution, collaborations, individual staff members, and students.
- Use appreciative inquiry approaches rather than deficit models of engagement.
- Consider embedding principles of assessment for learning in University policies and procedures.
- See students as active partners and contributors to all phases of the project.
- Build on and learn from previous centrally led and locally led faculty-based assessment and feedback initiatives.
- Be inclusive and develop assessment leaders across the University.
- Build and cascade capacity in assessment expertise.
- Use and develop governance of education, including reporting and action planning around assessment. Ensure that associated systems, workflows and regulations help assessment endeavours. Ensure that creativity flourishes in assessment.
- Create ways of working with clear communication channels and lines of responsibility and ownership.
- Align to the mission and strategy of the University.
- Become research-informed and evaluative in approach.
- Enhance educational effectiveness and resource efficiency of assessment activities.

9.3.2 Project Governance

The importance of this work is demonstrated by the strategic oversight from the University's Vice Principal (Education).

There was a significant intent in the project to become collaborative, collegiate, and support the development of assessment literacies for both staff and students. The project sought to secure a long-lasting and positive impact rather than promote short-term responses and so called quick-wins. To assist with this objective, the project established four responsibility areas:

- *Institutional responsibility*: systems and processes assure academic quality without stifling creativity in the assessment domain.
- *Team responsibility*: ensuring collective responsibility to an assessment agenda within schools, departments and teams.
- *Individual responsibility*: ensuring that staff who engage students understand students' personal importance and the impact of their assessment activity.
- *Student responsibility*: Developing agency and assessment literacy.

In the first year, the central project team was working with programme teams as pilots. Pilot programmes were drawn from across the University:

- Bachelor of Laws (LLB Undergraduate programme).
- Pre-Registration Nursing (BSc Undergraduate programme).
- Management (BSc) (Undergraduate programme).
- English (BA) (Undergraduate programme).
- MSc Mental Health (Postgraduate programme).
- MSc Cognitive Behavioural Therapy (Postgraduate programme).
- MSc Genes and Gene Therapy (Postgraduate programme).

The programmes were identified an Assessment Leader where such a role does not already exist, in order to provide leadership within the Faculty and a single point of contact. Enthusiasm to engage at the early stages is important, as these 'early adopters' will become 'champions' within each Faculty and, working with the Assessment Leaders, will help bring about a process of wider-development and support a response to any locally prevailing assessment concerns. The project structure is outlined in Fig. 9.2. The targeted areas were supported for inter-faculty diffusion of experience and development.

In all responsibility areas, the project's organisation refers to other relevant parts of the University. Overarching the ways of working is an appreciative inquiry approach. It sets out an institutional direction whilst allowing Programme teams to account for local programme and subject specialties. Russell et al. (2013) have shown repeatedly that an appreciative inquiry approach is more helpful because it

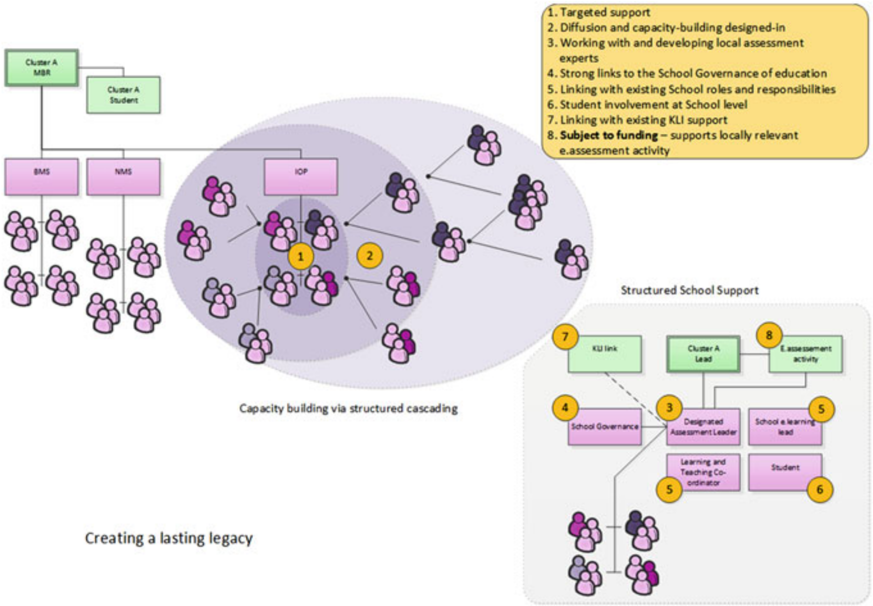

1. Targeted support
2. Diffusion and capacity-building designed-in
3. Working with and developing local assessment experts
4. Strong links to the School Governance of education
5. Linking with existing School roles and responsibilities
6. Student involvement at School level
7. Linking with existing KLI support
8. **Subject to funding** – supports locally relevant e.assessment activity
Cluster A MBR
Cluster A Student
BMS
NMS
IOP
1
2
Capacity building via structured cascading
Structured School Support
7
KLI link
Cluster A Lead
8
E.assessement activity
4
School Governance
3
Designated Assessment Leader
School e.learning lead
5
Learning and Teaching Co-ordinator
5
Student
6
Creating a lasting legacy

Fig. 9.2 An overview of the project structure

illuminates existing success rather than highlights deficits. For those readers looking for challenges to resolve, programme teams often highlight issues or concerns even when a discussion focuses on things going well. It takes a strong and observant leader to bring the group back to focusing on strengths.

The broad activity in the pilot includes:

- Establish the current status. Visually map the assessment landscape of the pilot programmes. The mapping activity includes:
  - – timeline of assessment (on the modules of the degree programme),
  - – stakes of the assessment (low, medium, or high in relation to marks),
  - – nature of the assessment activity (essay, examination, presentation, etc.) and,
  - – links between any assessment tasks (links within and across the modules).
- Establish the rationale for the current assessment strategy.
- Assessment redesign and making more use of existing effective and efficient assessment.
- Evaluation and diffusion.
- Operate in a continuous cycle of generative activity with continuous evaluation to generate new ideas and involvement by teams and faculty.

For future work, and to build opportunities for growing and sustaining the impact,

- Plan for and create an assessment review and redesign process to enable other programme teams to undertake assessment reviews and redesign.

A sample assessment landscape is shown in Fig. 9.3. The circles indicate the assessment activity (colour coded and scaled to indicate the weighting of the task),

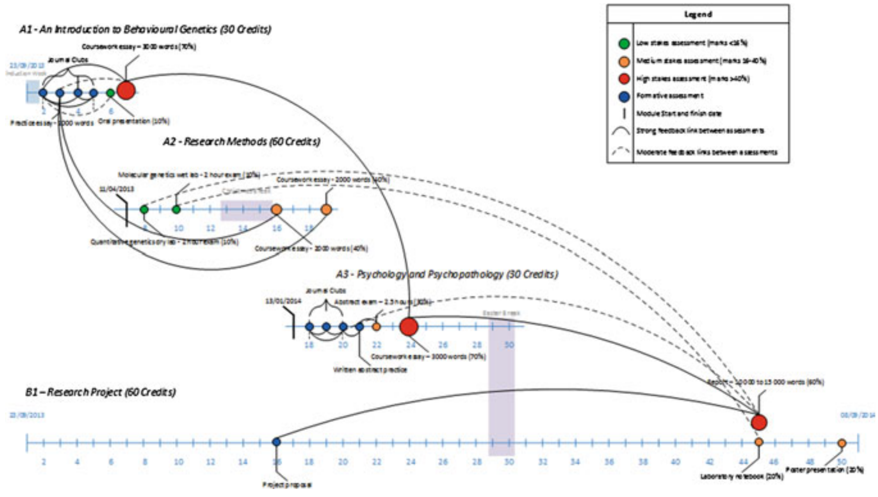

A1 - An Introduction to Behavioural Genetics (30 Credits)
Coursework essay – 3000 words (70%)
Journal Clubs
23/09/2013
Induction Week
2
4
6
Practice essay – 2000 words
Oral presentation (10%)
A2 - Research Methods (60 Credits)
Molecular genetics wet lab - 2 hour exam (10%)
Coursework essay - 2000 words (40%)
21/04/2013
Christmas break
8
10
12
14
16
18
Quantitative genetics dry lab - 2 hour exam (10%)
Coursework essay - 2000 words (40%)
A3 - Psychology and Psychopathology (30 Credits)
Journal Clubs
13/01/2014
Abstract exam – 2.5 hours (30%)
Easter Break
18
20
22
24
26
28
30
Coursework essay – 3000 words (70%)
Written abstract practice
B1 – Research Project (60 Credits)
23/09/2013
Report – 10 000 to 15 000 words (60%)
08/09/2014
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
Project proposal
Laboratory notebook (20%)
Poster presentation (20%)
Legend
Low stakes assessment (marks <15%)
Medium stakes assessment (marks 16-40%)
High stakes assessment (marks >40%)
Formative assessment
Module Start and finish date
Strong feedback link between assessments
Moderate feedback links between assessments

Fig. 9.3 An example of an assessment landscape

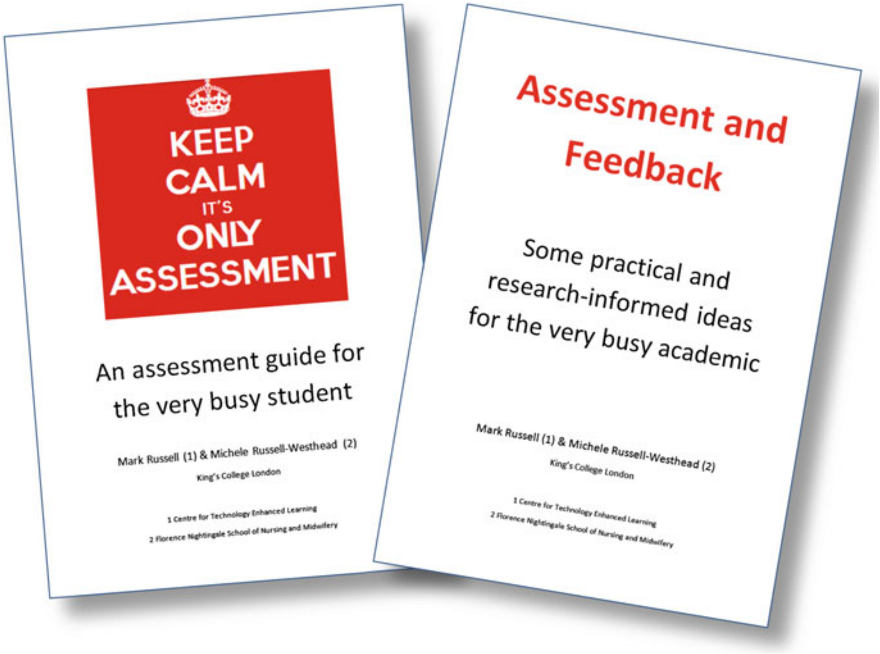
whereas the lines represent the learning-oriented links between the different assessment tasks. The recycling process is also indicated.

9.3.3 Emerging Findings

There are significant strengths in some of the assessment designs as well as opportunities to reconsider such assessment designs. Some specific and immediate emerging findings include:

- An interest in illuminating assessment activity at the programme level rather than at the module level,
- An emphasis on the high stakes end of process examinations,
- An emphasis on essays and examinations that are not administered in public,
- Challenges of creating an educative assessment experience as class size increases, and
- Challenges around consistency of marking.

Emerging findings show that faculty members and students need to become more confident and comfortable with assessment literacies. Two additional resources are being produced (see Fig. 9.4). Essentially, separate support guides that provide research-informed evidence to assist busy students and faculty members. The team



KEEP CALM IT'S ONLY ASSESSMENT

An assessment guide for the very busy student

Mark Russell (1) & Michele Russell-Westhead (2)
King's College London

1 Centre for Technology Enhanced Learning
2 Florence Nightingale School of Nursing and Midwifery

Assessment and Feedback

Some practical and research-informed ideas for the very busy academic

Mark Russell (1) & Michele Russell-Westhead (2)
King's College London

1 Centre for Technology Enhanced Learning
2 Florence Nightingale School of Nursing and Midwifery

Fig. 9.4 Assessment guides for students and faculty members

intends such resources, not to assist faculty members to become assessment researchers, but rather to offer informed advice in relation to the need, design, development, implementation and evaluation of their assessment endeavours.

9.4 An Illustrative Case Study: Strategic Development of Collaborative Assessment Practices in Nurse Education

The Florence Nightingale Faculty of Nursing and Midwifery at King's College London undertook a number of curriculum, assessment, and staff development initiatives. Their intention was to sharpen the focus on students' educational experiences and engage in supported faculty reflection and development. What follows is an overview of assessment initiatives and an illustrative case study of a faculty-wide response to some of the assessment challenges. The work is centred on a collaborative approach to designing and utilising marking rubrics to improve both teaching and learning.

9.4.1 The Challenge

As mentioned, the NSS and other student evaluations suggested improvements in assessment, marking, and feedback were necessary. The King's *Assessment and Feedback Project* afforded a timely, practical, and strategic opportunity to work with university partners to cohere our own evaluation and develop our own assessment practices.

The pre-registration BSc (Hons) Nursing programme, the largest undergraduate degree programme in the Faculty, was chosen as a pilot.

Additional data was gathered and analysed to establish and confirm the nature of some of the assessment challenges. Essentially, an audit exploring the quality of feedback and its compliance with procedures and alignment with good practice (Jackson and Barriball 2013) was undertaken. This audit helped clarify areas that required further guidance, support, and reiteration. In parallel, student focus groups examined their experiences of assessment processes and how feedback supported their learning. Faculty focus groups examined their experiences of the assessment process. From audit and focus group data, key issues clustered in three areas: Consistency, Transparency and Engagement:

Consistency

- Disparity of grades between markers.
- Variability and incongruence of feedback provided.

Transparency

- A lack of shared understanding of the criteria within and across module teams.

- Variability to the process brought about by non-faculty markers (clinical fellows).
- University assessment criteria were too generic and did not fully reflect variation across the levels or between grade bands effectively or accurately (For example: level 4 criteria not substantially different from level 6 and a C grade not clearly different from a B grade).

Engagement

- Ineffective communication of marking criteria to students.
- Little or now use of the criteria and guidelines as a means of improving student work.
- A need for students to take greater control of their learning.

These findings highlighted a need to:

- Share good practice in assessment and feedback;
- Support staff in developing educationally effective assessment and feedback;
- Engender more collegiality in relation to teaching, learning and assessment;
- Engage students more fully in the design, implementation and evaluation of learning and assessment practices.

9.4.2 The Response

The findings of the audit were used to inform improved assessment practices. One of these strategic activities was to introduce 'module-specific' marking rubrics to replace the generic university marking criteria. Given that lack of transparency and consistency of marking were key issues, it was decided that a bespoke rubric for each module would be more helpful to both staff and students.

A rubric is 'a grid of assessment criteria describing different levels of performance associated with clear grades' (Reddy and Andrade 2010, p. 435). It can be used for marking assignments, class participation, or overall grades. Research suggests that a module-specific rubric has major advantages. It provides a clear relationship between learning outcomes, learning activities and assessment (Hyland et al. 2006). The objectives of this project were to explore:

- The value in using marking [grading] rubrics.
- Evaluate student learning and assessment experiences.
- Establish the 'active ingredients' of embedding consistency, transparency and engagement into our assessment and feedback processes to improve and enhance nurse education.

An evaluation matrix was constructed to review the outcomes of the work (see Table 9.1).

Table 9.1 Evaluation metric

	Objective 1	Objective 2	Objective 3
Objectives	Explore the value in using marking rubrics to improve consistency and transparency in the assessment and feedback process	Engage students in the process of design, development, implementation and evaluation of their learning and assessment experiences	Establish the 'active ingredients' of embedding consistency, transparency and engagement into assessment and feedback processes to improve and enhance nurse education
Impact measures	Level of complaint around consistency and transparency of teaching and assessment reduces	Student voice is reflected in the design of module assessments, design of the rubrics and evaluation, of the impact on their learning	Student assessment feedback sheets reflect a shared language across modules and programme
			Best practice and student feedback is used to inform on-going decision making in education practices and curriculum documentation
	Rubric usage provides a effective and efficient means of marking and providing feedback		A range of formative assessment activities are activities are included in each module
Process indicators	A range of activities using the rubrics developed, used and evaluated with students	Students form a part of module team assessment design meetings, curriculum approvals and reviews, design of evaluation forms, frameworks, and handbooks	Regular module team meetings to discuss teaching and learning and co-construct assessment tasks, marking, and feedback processes a. Collaborative 2nd marking and moderation practices b. Creating of a bank of 'quick marks' in TurnitIn specific to each module c. Module-specific rubrics developed for all undergraduate nursing level 6 modules d. Mentoring of clinical and less experienced academic staff
	a. Case studies developed		
	b. Workshops on assessment and feedback, curriculum design, and teaching innovation		

Note Turnitin is an online originality checking tool that also includes online assessment functionality

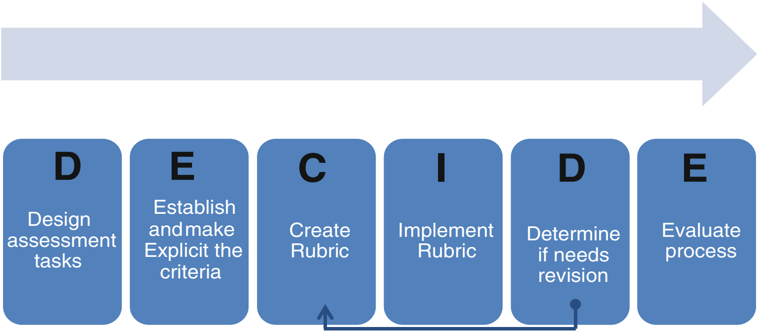


Fig. 9.5 DECIDE model

A series of workshops were developed and run on best practice in assessment. They included, assessing professional and clinical competence, creating assessment rubrics, and developing collaborative enhancement practices.

Four pilot modules were selected for the rubrics created. Two undergraduate pre-registration modules and two post-registration modules, using the DECIDE model (Russell-Westhead 2014). Essentially, the DECIDE model is a structured approach to the design, implementation and evaluation of a rubric (see Fig. 9.5).

Design

The module teams, and some students, met to decide how to develop rubrics that clearly linked to learning. Establish and make Explicit the key aims and outcomes of the assessment.

Create

The rubric was created by the team using a template based upon the University's generic marking criteria. It made explicit professional competences and application of knowledge to nursing practice (evidence-based practice) not reflected in the institutional mark scheme. Rubric descriptors became quick marks on *TurnItIn.* This used the language of the assessment guidelines and reflected the aims and outcomes explicitly at each level.

Implement

The rubrics were used in various assessment contexts which are briefly introduced below:

9.4.3 Module 1—Self-Assessment

Neonatal Care (Level 5 Module on the Midwifery BSc Programme)

All students were invited to bring their full first draft assignment to a writing workshop after the last session of the module. They were given the marking criteria (the assessment rubric), the assessment task and a grading sheet and asked to critically appraise their own work as if it were another student's work. They were to underline the comments in the rubric, which they felt best related to the essay and then provide a mark and feedback based on the assessment they had made. They then were asked to use the feedback to consider how they then could improve their work.

9.4.4 Module 2—Peer Assessment

Qualitative Research Methods (Level 7 Master's module)

The approach was similar to the self-assessment activity of Module 1, but this time the submissions were peer reviewed. Students were given another student's work to mark with the rubric, task and grade sheet. They were given 30 min to mark the work, assign a grade and write some developmental feedback. The tutor and researcher then monitored the student activity to ensure that everyone was engaging with the task correctly and fairly. The students were subsequently asked to discuss and justify their marking and comments to the person sitting next to them, which emulated the moderation process. This provided an opportunity to learn from other students' experience.

9.4.5 Module 3—Tutor Led Assessment

Research Project on all Undergraduate Programmes (level 6 pre-registration)

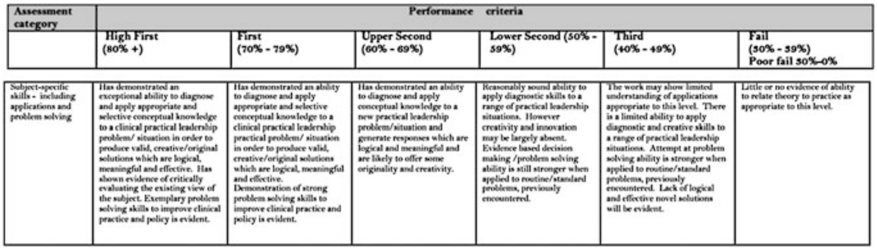
This activity used a number of exemplar assignments from the previous year in each of the grade bands. These exemplar assessments were assigned to groups of three students and they were to mark the assignment as if they were the second marker using the rubric and guidelines and provide feedback. They then discussed their marking decisions as a group to compare their marks and comments. A group discussion followed as to what made a good and less favourable assignment. The tutor provided group feedback and support in how to improve using meta-cognition to tackle that assignment.

The students were then encouraged to apply the same approach with their own work prior to submission of their assessment. Their individual project supervisors/tutors supported the process but it was not compulsory to use this approach.

9.4.6 Module 4—360 Review

Leading and Managing Care (Level 6 post-registration module)

It was decided by the module lead that the design of an assessment rubric provided an opportunity to completely rethink both the assessment task and the approach to teaching and learning throughout the module. There were a number of formative tasks throughout the course, such as presentations, critical evaluations of reports, writing



Assessment category	Performance criteria					
	High First (80% +)	First (70% - 79%)	Upper Second (60% - 69%)	Lower Second (50% - 59%)	Third (40% - 49%)	Fail (30% - 39%) Poor fail 30%–0%
Subject-specific skills - including applications and problem solving	Has demonstrated an exceptional ability to diagnose and apply appropriate and selective conceptual knowledge to a clinical practical leadership problem/ situation in order to produce valid, creative/original solutions which are logical, meaningful and effective. Has shown evidence of critically evaluating the existing view of the subject. Exemplary problem solving skills to improve clinical practice and policy is evident.	Has demonstrated an ability to diagnose and apply appropriate and selective conceptual knowledge to a clinical practical leadership practical problem/ situation in order to produce valid, creative/original solutions which are logical, meaningful and effective. Demonstration of strong problem solving skills to improve clinical practice and policy is evident.	Has demonstrated an ability to diagnose and apply conceptual knowledge to a new practical leadership problem/situation and generate responses which are logical and meaningful and are likely to offer some originality and creativity.	Reasonably sound ability to apply diagnostic skills to a range of practical leadership situations. However creativity and innovation may be largely absent. Evidence based decision making /problem solving ability is still stronger when applied to routine/standard problems, previously encountered.	The work may show limited understanding of applications appropriate to this level. There is a limited ability to apply diagnostic and creative skills to a range of practical leadership situations. Attempt at problem solving ability is stronger when applied to routine/standard problems, previously encountered. Lack of logical and effective novel solutions will be evident.	Little or no evidence of ability to relate theory to practice as appropriate to this level.

Fig. 9.6 Extract from rubric

strategies and observing work practice and being observed in their clinical role. These tasks were either self-assessed (reflection), peer-assessed (critical evaluation), or tutor (or work colleague) assessed (critical judgment). The purpose of doing so was that the participant could see their work from multiple perspectives to offer several opportunities to learn from others. With support and facilitation the summative task and the rubric were designed by the entire module team (7 lecturers) along with a small group ($n = 4$) of 'consultant students' who self-selected to get involved. They collaboratively decided upon both the assessment task and the criteria for grading the assignment (Fig. 9.6).

Determine the Appropriateness of the Criteria—the rubric was evaluated by the staff as to the value, usability, and transparency of the criteria and where necessary modifications made and finally.

9.4.7 Evaluation

A pragmatic qualitative approach (Creswell 2007) was adopted to describe student and staff views and experiences of the rubric project. All students completed a standard module survey ($n = 126$) with an additional question about the use of the rubrics and associated class activities. In addition, student focus group interviews were carried out at the end of each of the pilot module and six semi-structured one to one interviews. The audio from the interviews were digitally recorded and transcribed. The survey data, transcriptions and interview notes were uploaded into QSR's NVivo 9 qualitative data analysis software.

Additional triangulation involved examination of the module handbook and feedback sheets. The purpose of the document review was to address the overarching project impact measures. In particular, the measures are around consistency and transparency.

The survey data, focus group and one to one interviews were analysed using thematic analysis (Braun and Clarke 2006) and constant comparison techniques (Rapley 2010).

The data from students and staff were analysed and responses for each group were compared within each theme to identify any similarities and differences between them.

Data saturation had been reached as no new themes were emerging on data analysis of the last interview transcript. Vignettes or quotes are used to illustrate aspects of each theme arising from the interview questions presented in Table 9.2.

Table 9.2 Evaluation questions

Student survey and focus group interview questions	Staff interview questions
1. Where you involved with developing the 'assessment strategy' for this module. If so how so, and if not why not, would you like to have been? 2. Did the rubric help your understanding of the marking criteria/expectations of the module? Yes/No If so, in what way? 3. Were the rubrics and assessment guidelines explained to you? 4. Did you find the class activities using the rubric helpful? If so, in what way? 5. Do you think having a module specific rubric helped improve Your work? Your engagement with staff and other students? Your writing and/or other academic skills? If so, what? 6. Would you like to have a rubric for other modules? If so why? If not, why not? 7. Do you have any ideas for improvements or activities to engage students more with assessment and feedback?	1. How was the assessment strategy and tasks designed for this module? How involved were both staff and students in the process? 2. Did the rubric help your understanding of the marking criteria/expectations of the module? Yes/No If so, in what way? 3. How did you use the rubric in class with the student? Please explain 4. Do you think this helped improve students work engagement with staff and other students academic skills 5. Would you use rubrics again? If so why? If not? 6. Do you have any ideas for improvements or activities to engage students more with assessment and feedback?

9.5 Findings

The key findings of the evaluation are:

9.5.1 High Levels of Engagement

Engaging with the criteria

Formative tasks engage students with the marking criteria and introduce them to the process of self-reflection on their academic skills. Students commented very positively on how the focus on self-evaluation helped them critique and improve their work. They also claim that using rubrics helped them to focus their efforts, produce work of higher quality, earn better grades and feel less anxious about assignments:

> …the tutors went through the criteria with us and gave examples of what a good paper looked like. I used the rubric to make sure I had done what was asked of me and the self-

> assessment made me fairly confident I would get a B possibly an A. I actually got 74 % [A], my best grade on the course so far. (Module 1 [M1], Student 4[S4])

> The self-assessment activity made me really think about the quality of my work because it was easy to look at the rubric in each section and see if I met the criteria. It made me less stressed about handing it in because I felt confident that I had passed (M1, S6,)

Students on the research methods module (Module 3) used assess the on-going progress of their work. This was perhaps because of the more independent nature of a research project and the different layout of the rubric:

> I use it to plan each section of my research proposal to make sure I covered everything (M3, S18)

However, some students also claimed that they did not use the criteria. This may have been because: rubrics were offered as optional tools. Some individuals may have felt confident they were progressing satisfactorily and the extra work involved in using rubrics on top of their very heavy assessment load was a disincentive.

> I've done a degree already so didn't really use it (M3, S13)
> My supervisor was brilliant and explained things really well so I only used it at the end as I had a last minute panic that I'd forgot something (M3, S4)

All of the activities required the students to act as assessors and so they are taking a critical perspective on the tasks that they are marking. *Engaging with each other.* A particularly impactful feature of using peer assessment is that the students see how other students tackle the same task which has additional benefits which include them actively seeking clarification on the assessment criterion (they had not in Module 1 when they self-assessed) and checking their own work thoroughly to make sure their peer had not made a mistake.

> I learned a lot from seeing how the others got on with the task (M2, S1)
> I could hear what other people were saying which helped me think about what I was saying about the work I marked- not just I liked it but this relates to the criteria because… (M2, S9)

In the peer assessment task (Module 2), essays were handed back into the lecturer and returned to their owner to review the comments against the criteria and reflect on whether or not they felt the grade assigned was fair. Most of the students then started checking the accuracy of their mark and feedback for themselves:

> When I got my essay back I remarked it myself to make sure they hadn't made a mistake and given me a lower grade (M2, S2)

They considered it to be high stakes because they were marking someone else's work thus were more rigorous in the task.

> I felt under a lot of pressure to be fair and give the other student's work an accurate mark and good feedback because someone else was doing the same with my work. (M2, S10)

One student also offered some appreciation for the role of the tutor in the assessment process:

> I had no idea how hard it was to mark an essay and how long it took (M2, S3)

All of this appeared to help them to develop their reflective capacity, self-awareness and deeper understanding of the assessment process and its role in learning.

9.5.2 Improvement in Consistency of Grades and Feedback

Students in all modules also commented on the levels of consistency of the amount and quality of feedback across the cohort, although there were mixed views on this:

> I liked the fact we got lots of feedback telling us what we had done well or not so well in the essay and as this was the same descriptions [criteria] that we'd been using in the rubric. There was more what the tutor calls 'feed forward' than I normally get but as it is the last module I can't really learn from it. (M3, S14)

> My feedback was the same as other peoples, so I don't even though if the tutor actually read it. (M3, S1)

The sentiment of the second comment was also seen in some shape or form in all of the other module examples. The students had asked for 'consistency' of feedback, language but when that aspect was addressed the new complaint was that they were '*getting the same comments as [their] friends*'. There was also some variation reported in the volume and usefulness of the feed forward section. Some tutors took the opportunity to provide specific comments on the quality and creativity of the work for students attaining the higher grades providing positive encouragement to publish or go on to a higher degree, which was widely appreciated by the students who received it. This was also picked up in the documentation review across the module examples. This was a marked improvement from the original audit findings, which showed a lack of written commentary and feed forward for high performing students.

Additionally, the documentation review revealed high levels of consistency between markers in amount of feedback and quality of feedback for individual student development. The grade given and the feedback provided were also consistent, i.e. the amount and quality of the written comments reflected the development required of the student to improve.

9.5.3 Empowerment

Ultimately, the combination of the rubrics in themselves and a more collegiate and engaging process appeared to have led to a sense of empowerment in the students over their own learning and their wider university experience. The students that were involved in the collaborative construction of the task and rubric provided them with a say in the choice of topic, method, criteria, weighting and timing of assessments which provided a sense of ownership, they described the experience as like:

> being in the 'circle of trust' like Ben Stiller in 'Meet the Fockers' (M4, SC1)

> getting a insider tip on a horse race or football match' (M4, SC3)

The overall feeling appears to be one of belonging, being 'part of the gang', feeling important, valued, listened to, and also the 'insider tip' comment was suggestive of that engaging more got you a better mark although it is unclear as to why they felt that. One of the student consultants, however, drew attention to the power relations at play despite the faculty's best efforts:

> I felt a bit awkward, like I didn't really belong there. It didn't seem right that the other students were making suggestions to the staff- seemed a bit cocky really, so I didn't really say anything… I suppose I was worried that if they didn't like what I had to say it might affect my mark [grade]. (M4, CS2)

The module evaluations revealed somewhat of a dichotomy through. Some students ($n = 4$) who weren't involved in the design phase stated that they felt '*disadvantaged*' because they hadn't been involved. Others commented that they wish they had been more involved as the student consultants '*seemed more confident*', although others were more disparaging suggesting that '*they were matey [friends with] with the lecturers*' and '*always took the lead giving nobody else a chance*'. It is unclear (and unmeasured) as to whether it was the engagement in the design team that gave them the confidence or leadership role (positively or negatively) or whether it was those characteristics that made them volunteer in the first place.

9.5.4 Discussion and Conclusions—Creating Rubrics… but oh so Much More

The overarching aim of this illustrative case study was to improve consistency and transparency of the assessment, marking and feedback process and for the most part, this has been achieved. However, the secondary but arguably more impactful outcome has been more active engagement of staff and students in the assessment process resulting in feelings of empowerment in both.

The findings indicated that the students in all Module Examples perceived the rubrics to be comprehensive, and linked to the specific assignment questions.

Students used rubrics in a variety of ways to ensure that they had met the assessment criteria and to improve their work as instructed by the tutor and some also used them as a guide to structure and assess the progress of their work.

Nearly all of the activities provided students with the opportunity to gain immediate and detailed feedback unique to that particular module.

Interviews and anecdotal evidence from conversations after the faculty development workshops indicated that the faculty collaborative approach to developing rubrics meant that all academic staff had a consistent message, were using the DECIDE model to shape their work.

Module teams had ensured greater levels of consistency and quality in both the grading and feedback provided to the students.

It was, however, the approach to marking and moderation that really made a difference in the consistency and justification of marking process. This has resulted in a recommendation to the Faculty Senior Team that time be allocated in the staff workload model for team-based assessment, i.e. to diffuse more widely the benefits gained from this work into other modules.

Module-specific rubrics proved the key active ingredient to ensure transparency in assessment criteria. The range of different approaches to explain rubrics used across the modules suggests that it is not how this is done, but that it is done that is of importance.

(a) The rubrics were introduced to the students at the start (with the above explanation), in the middle (as an activity) and towards the end before they carried out their assignments. This means that the students could use the rubrics as an aid to planning, writing their assignments and assessing their performance. This approach appears to underpin their claims that the rubrics have helped them focus their efforts and produce work of higher quality, but additionally feel much more prepared and less anxious about assignments.

(b) It is necessary to have some discussion with the group about the marks and provide some quality assurance that the marking criteria had been fairly applied. For this reason, it is recommended that it is an anonymous activity (i.e. the papers are blind-marked so that the student cannot be identified) and have academic staff available for questions and monitoring

(c) Having the assessment information in written format in the module handbook and digitally on TurnitIn demonstrated both transparency and consistency and minimised the potential for conflicting advice between members of staff, students not having to rely on their own notes and interpretations of the verbal explanations provided which potentially could lead to misunderstanding.

The Assessment and Feedback Project succeeded in making the academic world transparent by inviting all staff to the assessment development meetings. In, some modules students and stakeholders actively participated in the project design and activity. All staff and students who provided feedback talked about having a sense of ownership and how empowering that is. Even staff and students who did not actively get involved in the design phase were pleased that they had been asked and several commented that they now wished they had been more involved.

In-class activities, empowerment came from sharing the assessment criteria (transparency) and being taught how to use it to improve (engagement) with the learning process. Students grew in confidence about the assessment judgments they were making. They used their knowledge and the skills as assessor, which developed their academic skills, professional skills and attributes. Building the capacity to make assessment judgments needs to be an overt part of any curriculum and one that needs to be fostered (Boud and Falchikov 2007).

9.6 Conclusion

The illuminative case study within the Florence Nightingale Faculty of Nursing and Midwifery was a reminder that teachers, their students and clinical teaching partners need the conditions to develop their own approach to improving feedback and assessment. However, their assessment and feedback project was initiated, informed by and reflected the values and priorities of a university wide strategy and project. In essence, it was a faculty project created to respond to their own demonstrable challenges (and aspirations) that was prioritised and enabled by the university.

The University's focus on feedback and assessment was externally driven. It was in part a response to policy and policy instruments. In this example, the need for change was from persistent external pressure, (NSS results and the associated league tables and heightened student expectations), but the change itself came from within. The model of change here is not one of top-down change but rather a systematic identification of a need, seen from the perspective of range of stakeholders, that is examined and addressed at local level. This model is owned, governed and makes sense to the staff in the faculty. It respects context and the importance of role of faculty and students in driving practice improvements in learning and assessment.

Issues/Questions for Reflection

- Higher educational institutions need to pay close attention to prevailing and emerging governmental policies to anticipate how they may help develop and/or challenge educational practices. What processes are there within your institution to do this?
- To what extent does your institution's educational strategies and practices respond to government policy and with what effects?
- This chapter provides evidence that an appreciative evaluative approach enables broad engagement in change and a more nuanced response to improving assessment practices. How are programme teams and academic departments engaged in curriculum planning and curriculum review discussions in your context? Do the discussions adopt a deficit evaluation of a current situation, (i.e. what is not working and how do we fix it?) or an appreciative inquiry approach used (i.e. what is currently working well, and how can we do more of that)?

Appendix

King's College London is the fourth oldest university in England. Founded in 1829 by King George IV and the Duke of Wellington, it was one of the two founding colleges of the University of London in 1836. King's has a world-renowned reputation in a number of disciplinary areas and ranks highly in international and national university league tables.

Organised in eight academic Faculties, King's has research-intensive orientation and is hence a member of the UK Russell Group. By their own definition, the UK Russell Group represents 24 leading UK universities which are committed to maintaining the very best research, an outstanding teaching and learning experience, and unrivalled links with business and the public sector.

Approximately 25,000 students study at King's, the majority of which study on a full time basis. Around 15,000 students are registered on undergraduate programmes, 7500 are registered on postgraduate taught programmes and the remaining (circa) 2500 are registered on postgraduate research programmes. Having a significant research-intensive history and culture is not without its challenges. King's is facing up to these challenges and is continually seeking to enhance the educational experience of its students. The College's research culture can also benefit its students and society.

References

Biggs, J. (2003). *Teaching for quality learning at university* (2nd ed.). Berkshire, NY: Open University Press.

Boud, D., & Falchikov, N. (2007). Rethinking assessment in higher education: Learning for the longer term. Abingdon: Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches*. Thousand Oaks, CA: SAGE.

Garnham, N. (2002). Information Society's theory or ideology: A critical perspective on technology, education and employment in the information age. In W. H. Dutton & B. D. Loader (Eds.), *Digital academe. The new media and institutions of higher education and learning. London.* Dordrecht, Heidelberg, London, New York: Routledge, Springer.

Hammond, N. (2007). Preface. In A. Jenkins & M. Healey (Eds.), *Linking teaching and research in disciplines and departments* (p. 3). Retrieved September 10, 2008, from (2012) Academic identities and research-informed learning and teaching: issues in higher education in The United Kingdom (p. 236) http://www.shapeoftraining.co.uk/reviewsofar/1788.aspin

Jackson, C., & Barriball, L. (2013). *Audit of assessment practices in the Florence Nightingale School of Nursing and Midwifery* (unpublished paper).

McKee, A. (2012). Academic identities and research-informed learning and teaching: Issues in higher education in The United Kingdom (p. 236). http://www.shapeoftraining.co.uk/reviewsofar/1788.aspin. In A. Mc Kee & M. Eraut (Eds.), *Learning trajectories, innovation and identity for professional development*. Dordrecht, Heidelberg, London, New York: Springer.

Merry, S., Price, M., Carless D., & Taras, M. (2013). *Reconceptualising feedback in higher education: Developing dialogue with students.* London, New York: Routledge.

Price, M., Handley, K., O'Donovan, B., Rust, C., & Miller, J. (2013). Assessment feedback: An agenda for change. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students*. London: Routledge.

Rapley, T. (2010). Some pragmatics of qualitative data analysis. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 273–290). London, UK: Sage Publications.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435–448.

Russell, M. B., Bygate, G., & Barefoot, H. (2013). Making learning oriented assessment the experience of all our students. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students*. London: Routledge.

Russell-Westhead (2014). The Decide Model. A presentation to the Florence Nightingale Faculty of Nursing and Midwifery.

Further Readings

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education, 38*(8), 941–956.

Bovill, C., Aitkin, G., Hutchinson, J., Morrison, F., Scott, A., & Sotannde, S. (2010). Experiences of learning through collaborative evaluation from a masters programme in professional education. *International Journal for Academic Development, 15*(2), 143–145.

Campbell, F., Eland, J., Rumpus, A., & Shacklock, R. (2009). *Hearing the student voice: Involving students in curriculum design and development.* Retrieved June 1st 2011, from http://www2.napier.ac.uk/studentvoices/curriculum/download/StudentVoice2009_Final.pdf

Dunne, E., & Zandstra, R. (2011). *Students as change agents: New ways of engaging with learning and teaching in higher education*. Bristol: ESCalate/Higher Education Academy.

Greenaway, D. (2013). The future shape of training. http://www.shapeoftraining.co.uk/reviewsofar/1788.asp. Accessed July 6, 2015.

Healey, M., Flint, A., & Harrington, K. (2014). *Engagement through partnership: Students as partners in learning and teaching in higher education*. York: Higher Education Academy. Available at: https://www.heacademy.ac.uk/sites/default/files/resources/Engagement_through_partnership.pdf. Accessed April 14, 2015.

Healey, M., Mason O'Connor, K., & Broadfoot, P. (2010). Reflections on engaging students in the process and product of strategy development for learning, teaching, and assessment: An institutional case study. *International Journal for Academic Development, 15*(1), 19–32.

Health Education England-Kent, Surrey and Sussex. (2013). *The sound of the student and trainee voice*. http://kss.hee.nhs.uk/events/studentvoice/accessed, March 10, 2014.

Krause, K., & Armitage, L. (2014). *Student engagement to improve student retention and success: A synthesis of Australian literature*. York: Higher Education Academy. Available at: https://www.heacademy.ac.uk/sites/default/files/resources/Australian_student_engagement_lit_syn_2.pdf. Accessed April 14, 2015.

Krause, K. L., Hartley, R., James, R., & McInnis, C. (2005). *The first-year experience in Australian universities: Findings from a decade of national studies*. Melbourne: University of Melbourne, Centre for the Study of Higher Education.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. *Assessment & Evaluation in Higher Education, 38*(8), 919–940.

Panadero, E., & Dochy, F. (2014). Student self-assessment: Assessment, learning and empowerment. *Assessment & Evaluation in Higher Education, 39*(7), 895–897.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144.

Price, M., O'Donovan, B., Rust C., & Carroll, J. (2008). Brookes eJournal of learning and teaching: Promoting good practice in learning, teaching and assessment in higher education. *Assessment Standards: A Manifesto for Change, 2*(3), 1–2.

Thomas, L. (2012). *Building student engagement and belonging in higher education at a time of change: Final report from the what works? Student retention & success programme*. York: Higher Education Academy.

Thornton, R., & Chapman, H. (2000). Student voice in curriculum making. *Journal of Nursing Education, 39*(3), 124–132.

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education, 14*(3), 281–294.

Trowler, P. (2002). *The Future Shape of Training*. Final Report, commissioned by Health Education England (p. 55). Gildredge, Abingdon: Education Policy. http://www.shapeoftraining.co.uk/reviewsofar/1788.asp. Accessed on March 9, 2014.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: MIT Press.

Weller, S., & Kandiko, C. B. (2013). *Students as co-developers of learning and teaching: Conceptualising 'student voice' in professional development*. A paper presented at Society for Research into Higher Education, Celtic Manor, Wales (UK).

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge, UK: Cambridge University Press.

Westhead, M. (2012). Cultivating student voice to enhance the curriculum and influence institutional change: The case of the non-traditional student. In *6th Excellence in Teaching Conference Annual Proceedings* (pp. 59–74).

Wiggins, G. (1998). *Educative assessment*. San Francisco, CA: Jossey-Bass.

Yorke, M., & Longden, B. (2008). *The first-year experience of higher education in the UK*. York, UK: Higher Education Academy.

Chapter 10
Thinking Critically About the Quality of Critical Thinking Definitions and Measures

Lily Fountain

Abstract Leading policy and professional organizations are in agreement that critical thinking is a key competency for health professionals to function in our complex health care environment. There is far less clarity in professions education literature about what it really means or how to best measure it in providers and students. Therefore, in order to clarify terminology and describe the context, definitions, and measures in critical thinking research, a systematic review using the keywords "critical thinking" or the associated term "clinical reasoning," cross keywords nurse and physician, resulting in 43 studies is presented in this chapter. Results indicated that an explicit definition was not provid7ed in 43 % of the studies, 70 % ascribed equivalency to the terms critical thinking, clinical reasoning, clinical judgment, problem-solving, or decision-making, and 40 % used researcher-made study-specific measures. Full alignment of definition and measure was found in 47 % of studies, 42 % of studies reported validity for obtained scores, and reliability was documented in 54 % of studies. Keyword analysis identified six common constructs in nursing studies of critical thinking in nursing: individual interest, knowledge, relational reasoning, prioritization, inference, and evaluation, along with three contextual factors: patient assessment, caring, and environmental resource assessment. Recommendations for future research include increased use of purposeful samples spanning multiple levels of provider experience; documentation of effect sizes, means, SD, and n; adherence to official standards for reliability and validity requiring documentation of both previous and current datasets as appropriate; and greater use of factor analysis to validate scales or regression to test models. Finally, critical thinking research needs to use explicit definitions that are part and parcel of the measures that operationalize the construct, and alternative measures need to be developed that line up the attributes of the definition with the attributes measured in the instrument. Once we can accurately describe and measure critical thinking, we can better ensure that rising members of the professions can apply this vital competency to patient care.

L. Fountain (✉)
Faculty of Nursing, University of Maryland School of Nursing, Baltimore, MD, USA
e-mail: fountain@son.umaryland.edu

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_10

Takeaways

- Performance constructs such as critical thinking should be explicitly defined and measures should align with these definitions.
- Previous critical thinking research has usually examined learners at only one level of experience, such as students, new graduates, or experts. Future studies can be strengthened by examining learners at more than one level of experience.
- Domain-specific measures are preferable for performance assessment. For example, in nursing education studies, individual interest, knowledge, relational reasoning, prioritization, inference, and evaluation, along with 3 contextual factors, patient assessment, caring, and environmental resource assessment, were identified as common keywords.
- Professions education research is strengthened when AERA/APA/NCME standards for reliability and validity are followed.

10.1 Introduction

> Nurse Jennifer has been a maternity nurse for 5 years and has been fascinated by maternity nursing since her basic education in nursing. Today, she stares thoughtfully at her patient Mrs. Gablonsky. Nurse Jennifer sees something surprising. Mrs. Gablonsky's condition differs in a way the nurse does not expect for a woman who birthed a baby the previous day. Nurse Jennifer wonders about what is causing Mrs. Gablonsky's state and questions the patient closely to find out if there were any symptoms that could help explain Mrs. Gablonsky's condition. Nurse Jennifer compares Mrs. Gablonsky's condition to the other postpartum women she has treated in her career. She searches her mental data base for knowledge about complications that could be consistent with the symptom that surprised her. After a few moments of contemplation, Nurse Jennifer knows how to help her patient.

As this scenario illustrates, critical thinking using key cognitive processes is centrally involved in the quality of care provided by maternity nurses and other healthcare professionals. Leading policy and professional organizations such as the Institute of Medicine, Carnegie Foundation, and American Association of Colleges of Nursing are in agreement that critical thinking is a key competency for health professionals to function in our complex health care environment (American Association of Colleges of Nursing 2008; Institute of Medicine 2010; Cooke et al. 2010). There is far less clarity in professions education literature about what critical thinking and its analog in practice, *clinical reasoning* really mean or how to best measure it to ensure competence in providers and students. Toward this end, this systematic review was conducted to examine the quality of definitions and measures of critical thinking and clinical reasoning within the literature pertaining to health care professions.

What is meant by *critical thinking*? Facione (1990) gives one commonly used definition of critical thinking as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference" (p. 2). This present study also examined the term *clinical reasoning*, defined by Higgs and Jones (2000) as "the thinking and/or decision-making processes that are used in clinical practice" (p. 194). Although other terms have been used to describe clinical thinking, such as clinical judgment, problem-solving, and decision-making, the terms critical thinking (CT) and clinical reasoning (CR) were chosen as the basis for this systematic review because a preliminary electronic database search indicated they were the most commonly equated terms populating the recent research.

Specifically, from the 1980s to the present, critical thinking and clinical reasoning have been areas of intense research and clinical interest in nursing and medicine, as documented by several recent reviews of the literature (Brunt 2005; Chan 2013; Norman 2005; Ross et al. 2013; Simpson and Courtney 2002; Walsh and Seldomridge 2006a). Critical thinking has been identified as a key construct in core competencies for interprofessional collaborative practice and consensus statements on critical thinking in health professions education (Huang et al. 2014; Interprofessional Education Collaborative Expert Panel 2011). The most important reason for studying critical thinking in the health professions is that health care providers, educators, researchers, and policy-makers believe it leads to better health care. It has been argued that critical thinking has the potential to reduce morbidity and mortality for patients, and increase patient satisfaction with care. It can reduce health care costs by avoiding mistakes, unnecessary procedures, and unnecessary use of supplies (Benner et al. 2008; Kataoka-Yahiro and Saylor 1994).

However, problems with the definitions and measures used in critical thinking and clinical reasoning have been recognized (Krupat et al. 2011; Walsh and Seldomridge 2006a). Further, the quality of definitions and measures used in educational research on critical thinking and clinical reasoning affects the ability to evaluate which educational strategies are effective at promoting these skills (Brunt 2005). In addition, the context of health education research affects the quality of research (Ovretviet 2011). Merriam-Webster defines *context* as "the interrelated conditions in which something exists" (Context n.d.). The types of participant samples and research designs used in health education affect the ability of researchers to clearly define and measure constructs in the health professions (Waltz 2010).

In order to function with the same level of research rigor as clinical practice (Harden et al. 2000), clear definitions (Creswell 1994, 2014) and measures (Ratanawongsa et al. 2008) have been identified as essential steps in producing quality professions education research. With this increased focus on evidence-based practice during education and clinical practice, in combination with the other pressures on health care and education, it is vital that a shared base of terminology and psychometrically sound assessment tools be identified.

There is long-standing literature on the problems with definitions of CT terms, which fall into the categories of clarity, domain specificity, and equivalency of term usage. The lack of explicit, clear definitions of terms has been cited as a methodological issue by other researchers in educational psychology (Alexander et al. 2011). Alexander et al. (2011) found that many studies were not explicitly defining terms for higher order learning. Similarly, Brunt (2005) stated that "to evaluate CT, educators need a clear definition or framework to use" (p. 255), noting the deficit specifically for critical thinking in nursing studies. Carpenter and Doig (1988) also advocated for the importance of clear critical thinking definitions in order to measure it. Lack of domain specificity has also been cited as a problem with definitions of these constructs (Ennis 1991; Pintrich et al. 1991). Walsh and Seldomridge (2006b) concluded that domain-specific standardized critical thinking instruments produced mixed results in nursing studies due to domain-general definitions.

The tendency to equate different terms for critical thinking in the professions with each other has been noted by many others (Patel et al. 2004; Simmons 2010; Simpson and Courtney 2002). For one, Patel et al. (2004) noted that the terms *clinical reasoning*, *problem-solving*, and *decision-making* have been used in medicine to describe how physicians make decisions (p. 2). Simmons described decision-making, problem-solving, clinical judgment, and clinical reasoning as being used synonymously (p. 1152). Likewise, Simpson and Courtney (2002) stated that "the multiplicity of definitions of critical thinking proved to be a hindrance" (p. 7).

In addition to these problems of clarity of definitions of critical thinking and clinical reasoning, problems in the psychometric properties of measures of critical thinking have also been identified (Ennis 1989). The features of measures that have been examined in studies of methodological quality include alignment (Simmons 2010), reliability, validity (Cook and Beckman 2006), and commonality (Abrami et al. 2008). For example, alignment, the degree of match between a construct's definition with operationalized measures in a study, has been examined for self-regulated learning in education (Dinsmore et al. 2008). Ratanawongsa et al. (2008) noted that "educators should consider whether previously developed instruments are valid for their targeted objectives."

Concerns regarding reliability and validity of data obtained from measures of critical thinking have been identified; Cook and Beckman (2006), for example, criticize the practice of citing evidence from a previous study only, in order to support validity. Watson et al. (2002) found that the lack of agreement on definitions contributed to the lack of validity of resulting data (p. 423). The *Standards for Educational and Psychological Testing* (AERA 1999) recommend that researchers document reliability and validity.

The final area used to examine the quality of measurement tools is commonality, used here to describe the degree to which a measure is shared in use by other researchers. This was coded in a meta-analysis of critical thinking instructional interventions: Abrami et al. (2005) categorized the measures used as: standardized tests; tests developed by a teacher; tests developed by a researcher; tests developed by teacher–researchers; and secondary source measures, which are measures adopted or adapted from other sources. One of the reasons that type of measures has been examined in critical thinking research is that standardized test scores in particular did not show expected improvements in critical thinking over the course of professional education (Walsh and Seldomridge 2006a).

Thus, the need exists for a detailed and systematic description of the types and quality of definitions and measures of critical thinking and clinical reasoning used in the health professions. This study seeks to address that gap. The aim of this review is to lay the foundation for improving the conceptualization and operationalization of critical thinking and clinical reasoning in the health professions by addressing the following research questions:

1. What is the nature of the context in which critical thinking and its analog clinical reasoning have been examined?
2. How and how well are definitions for critical thinking and clinical reasoning specified in the literature?
3. How and how well are measures of critical thinking and clinical reasoning used in the literature?

10.2 Methods

10.2.1 Data Sources

I conducted a systematic review of the literature, using specific terms, delimiters, and selection criteria. A combination of electronic and hand searching were used to minimize bias in article selection (Cooper et al. 2009).

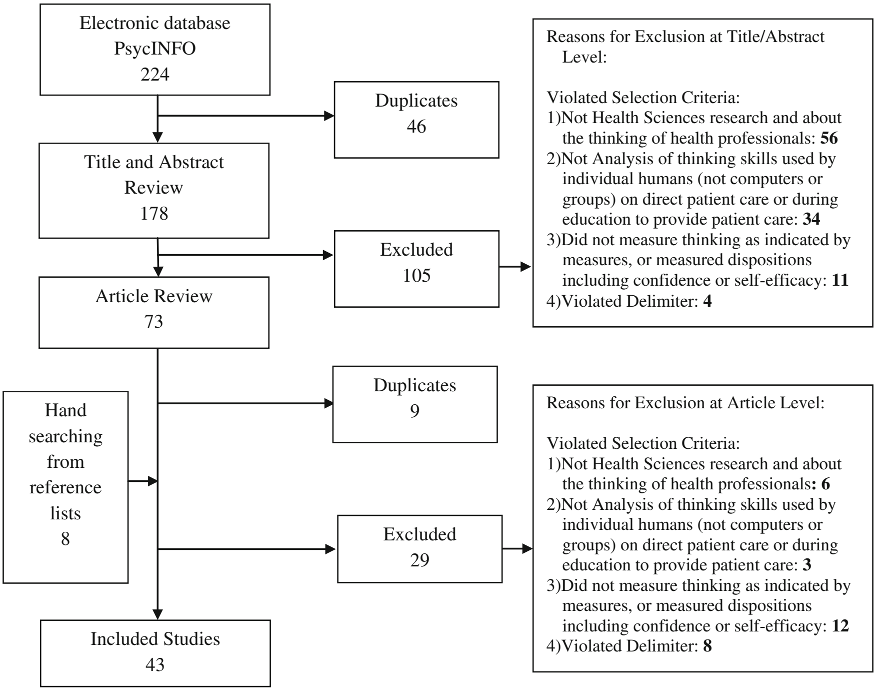
10.2.2 Study Selection

Inclusion criteria were peer-reviewed, empirical journal articles published between 2006 and 2011 using the terms clinical reasoning or critical thinking, searching the PsycINFO database. PsycINFO, as an interdisciplinary database for behavioral and social sciences research, that includes psychology, nursing, and education, was chosen as the database for this study in order to more closely examine the educational psychology research on this topic and maintain an educational focus, not a clinical one. Delimiters included the database: PsycINFO; keywords: clinical reasoning or critical thinking; cross keywords: nurs* and doctor, physician; language: English; type of article: empirical research; publication time period: June 2006 to June 2011; population: human; and publication type: peer-reviewed journals. To focus on the evaluation of critical thinking in practice, selection criteria required that included articles (a) studied direct patient care, or education to provide patient care, in the health sciences and (b) concerned the thinking of individual human health professionals (not computers), and (c) that the research directly measured thinking, not dispositions, confidence, or self-efficacy. The search strategy is summarized in Table 10.1.

In addition to the electronic search, footnote chasing was used to identify relevant studies in the reference lists of the study pool of articles. Eight articles that met the criteria that were identified by this method were included. The tables of contents of the electronic search's most commonly identified journal, *Journal of Advanced Nursing*, was also physically searched for the year 2011, but no new articles were identified. The resulting pool of articles comprised the data source for this systematic review. Figure 10.1 summarizes the disposition of articles in the

Table 10.1 Search strategies for literature

1. EBSCO online research platform
2. PsycINFO Database
3. Advanced string search—Boolean/Phrase
4. Critical thinking in title (or)
5. Critical thinking in abstract (or)
6. Clinical reasoning in title (or)
7. Clinical reasoning in abstract (and)
8. Nurs* (or)
9. Physician (or)
10. Doctor
11. Narrow results by source type: empirical periodicals
12. Narrow results by year (2006–2011)



Reasons for Exclusion at Title/Abstract Level:

Violated Selection Criteria:
1)Not Health Sciences research and about the thinking of health professionals: **56**
2)Not Analysis of thinking skills used by individual humans (not computers or groups) on direct patient care or during education to provide patient care: **34**
3)Did not measure thinking as indicated by measures, or measured dispositions including confidence or self-efficacy: **11**
4)Violated Delimiter: **4**

Reasons for Exclusion at Article Level:

Violated Selection Criteria:
1)Not Health Sciences research and about the thinking of health professionals**: 6**
2)Not Analysis of thinking skills used by individual humans (not computers or groups) on direct patient care or during education to provide patient care: **3**
3)Did not measure thinking as indicated by measures, or measured dispositions including confidence or self-efficacy: **12**
4)Violated Delimiter: **8**

Fig. 10.1 Summary of literature search and review process for primary literature

review process. As the figure indicates, 224 abstracts were produced by the search terms; after title, abstract, and full article review, 43 articles met the criteria for inclusion in the review.

10.2.3 Coding Protocol

In order to clarify the definitions and measures used for critical thinking research, an explicit coding scheme was used. I developed a protocol based on the coding typology used in prior research (Alexander and Murphy 2000; Dinsmore et al. 2008), and the recommendations of the Best Evidence Medical Education (BEME) collaboration (Harden et al. 2000). The coding typology for Dinsmore et al. (2008) was adapted to classify explicitness or clarity of definitions, and the alignment of measures with the definitions. The categories of study design, study size, target outcomes, and level of experience of participants were chosen from the BEME categories to evaluate the quality of medical education research studies.

Overall, the study variables were categorized as relating to the contextual aspects of the studies, the definitions, or the measures. A general description of the coding is given here, and Appendix A details the components of the codebook that was developed, that specified all resulting codes and was used to establish interrater agreement on the coding scheme.

Context. Variables coded for context include purpose, participant variables, and research design.

Purpose. The constructs critical thinking and clinical reasoning were examined for several purposes. McCartney et al. (2006) caution that care must be exercised when a measure is used differently than its intended purpose, such as either clinical evaluation, screening, or research. In this study, the purposes were categorized into four types: 1—evaluation of a program of education; 2—evaluation or description of a teaching technique; 3—evaluation of admission, course performance, or progression decisions; 4—evaluation or description of students, faculty, or providers.

Participants. The participants in the study pool were coded by professional *domain*, *level of experience*, and *number*. For this study, professional domain was determined at the level of licensure, such as nursing or medicine; the study may have focused on a subspecialty of the domain, such as emergency nursing or cardiac medicine but these were coded according to the overreaching domain category. In this study, medicine refers to the profession of doctors or physicians. Nursing refers to the profession of nurses, and includes registered nurses at all education levels. In addition to medicine and nursing, articles from veterinary medicine, kinesiology, health sciences, and occupational therapy were produced by the search. Since the goal for this study was an examination of terms used in critical thinking across multiple health care domains, these articles, which included references to medicine or nursing, were retained if they met the delimiters and selection criteria. In addition, two articles examined more than one profession, and were labeled *multidisciplinary*. Each study was coded for the level of experience of participants, either

(a) student or prelicensure, (b) new graduate or residents, (c) practicing provider, or (d) multiple levels. The number of participants or sample sizes were categorized as small, with less than 30 participants, moderate, with 31 to 100 participants, or large, with over 100 participants.

Research designs. The studies were categorized by type of research design. Preexperimental designs included one group pretest/posttest and cross-sectional studies. Quasi-experimental designs included separate sample pretest/posttest design, and separate sample pretest/posttest control design. Examples of experimental designs include pretest/posttest control group design and Solomon four-group design. Case studies were coded as qualitative.

10.2.4 Definitions

For each study, the definition or descriptive data about of critical thinking or clinical reasoning was coded for clarity, domain specificity, and equivalency.

Clarity. In this study, *clarity* refers to whether the definition was *explicitly* or *implicitly* defined in the study. A definition was coded as *explicit* if the author explicitly stated the definition of critical thinking used in the study. For example, in Funkesson et al. (2007), the following definition for clinical reasoning was explicitly stated: "In this paper, clinical reasoning is seen as a cognitive process, where both theoretical knowledge and personal experience are used in a unique care situation aiming to achieve a desired outcome for the person in focus" (p. 1110).

In order to analyze the lack of distinction in CT/CR term usage, the explicit category was further delineated. If the definition was explicitly stated, the definition was analyzed as to whether it was a definition *shared* by other researchers in published research, or an *idiosyncratic* definition used by the researcher for this specific study. For example, Blondy (2011) stated this definition: "We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation and inference… and inductive and deductive reasoning" (p. 182), and this was coded as explicit shared. Appendix B contains a list of shared definitions.

Forneris and Peden-McAlpine (2007), on the other hand, explicitly stated her own unique definition: "Grounded in these theoretical views, critical thinking is defined as a process of reflective thinking that goes beyond logical reasoning to evaluate the rationality and justification for actions within context…Using the work of these theorists, four attributes of critical thinking: reflection, dialog, context, and time (p. 411). This definition was coded as explicit idiosyncratic. *Idiosyncratic* explicit definitions were definitions that contained components that were specific to this study, not captured by previously published definitions.

Another example of an explicit but idiosyncratic definition is "critical thinking: problem identification, problem definition, exploration, applicability, and integration" (Schnell and Kaufman 2009). The keywords "applicability" and "integration" were not found in common CT definitions, so this study was coded as explicit idiosyncratic.

If the definition was only *implicitly* defined, the definitional data were further analyzed as to the manner in which the construct was discussed and was coded as conceptual, referential, or measurement. If the construct was discussed through the use of related concepts, it was coded as *implicit conceptual.* For example, Mamede et al. (2007) stated:

> Judicious judgements [sic] and effective decision making define successful clinical problem solving. Two different approaches for processing clinical cases, nonanalytical and analytical, have been shown to underlie diagnostic decisions. Experienced doctors diagnose routine problems essentially by recognizing similarities between the actual case and examples of previous patients. This pattern-recognition, non- analytical form of clinical reasoning is largely automatic and unconscious. In the second, analytical form of case processing, clinicians arrive at a diagnosis by analyzing signs and symptoms, relying on biomedical knowledge when necessary. (p. 1185)

Mamede's discussion used many concepts that are part of the discussion of clinical reasoning but no definition was stated. Thus, this definition was coded as implicit conceptual.

If the author did not clarify which definition was being used and cited definitions used in other literature, the definition was coded as *implicit referential.* For example, Göransson et al. (2007) stated that, "deductive content analysis was followed, using the thinking strategies *described by Fonteyn and Cahill* (1998) from a long-term TA study" (emphasis added). If the author only defined the construct through the use of the measure, the clarity was coded as *implicit measurement.* For example, Wolpaw et al. (2009) measured the following outcomes: summarizing patient findings; providing a differential diagnosis; analyzing possibilities in differential diagnosis; expressing uncertainties and obtaining clarification; discussing patient management; and identifying case-related topics for further study. Because they did not define clinical reasoning, but measured these aspects of clinical reasoning, so this study was coded as implicit measurement.

Domain specificity. Domain specificity refers to whether the definition used was specific to a domain or was generally applicable. In this study, if a definition was specific to a domain, it was coded as *domain-specific.* By contrast, if the definition could be used by other professions it was coded *domain-general.* For instance, Johansson (2009) defined clinical reasoning as "the cognitive processes and strategies that nurses use to understand the significance of patient data, to identify and diagnose actual or potential patient problems, to make clinical decisions to assist in problem resolution and to achieve positive patient outcomes" (p. 3367). The specification of the definition as pertinent to nurses rendered this a domain-specific definition. On the other hand, Ajjawi and Higgs (2008) stated "clinical reasoning is defined as the thinking and decision-making processes associated with professional practice," and this was coded as domain-general.

Equivalency. *Equivalency* refers to how the numerous terms used to describe critical thinking were used in the study. Up to 43 terms for critical thinking have been identified (Turner 2005). Often authors would state that one term was equivalent to another by using such phrases as “also known as” and “or.” This intermingling of terms was analyzed and coded as “equivalency.” For example, Funkesson et al. (2007) stated “clinical reasoning…can be named critical thinking, reflective reasoning, diagnostic reasoning, decision making, etc.” (p. 1110). This was coded as equivalency present. For the purposes of this study, the terms analyzed for statements of equivalency were critical thinking, clinical reasoning, clinical judgment, decision-making, and problem-solving.

10.2.5 Measures

In addition to definitions, the instruments or measures used in studies during this time period were coded for definition—measure *alignment, reliability, validity*, and *commonality* of measures used.

Alignment. Alignment of the definition under discussion with the measure/instrument used is valued in educational research. I used the typology described in Dinsmore et al. (2008) to categorize the amount of congruence of this operationalization of the definition and the measure for each article in this review. Only articles with explicit or implicit through concepts definitions were analyzed for alignment. Categories were defined as full (clear match between definition and measure), partial (partial match; only some of the items measured by the instrument were in the definition), and not applicable (referential or measurement definitions).

For example, per the instrument development description, the components of the California Critical Thinking Skills Test are based on Facione (1990) definition of critical thinking, so there was full alignment of definition and measure. By comparison, the study by Nikopoulou-Smyrni and Nikopoulos (2007) used the Anadysis Critical Reasoning Model to measure CR. The definition of CR stated that treatment should be “in collaboration with a patient or family (if possible).” However, the steps in the model (i.e., gather and evaluate data, predict risk, develop treatment plan, monitor progress) did not explicitly include the patient input in measurement so the study was coded as partial for alignment of definition with measure.

Keywords. In order search for common ground among the many definitions, the attributes described in definitions were summarized in key words or phrases, and a list compiled (see Appendix C). Keywords were chosen from the actual words used in the definition or a close synonym or a word or phrase given equivalence in the studies. For example, many studies used some phrase including *data*, such as *data collection*, *data gathering*, or *noticing*. *Significance* of data was a different keyword. If the definition was an explicit shared definition, the keywords were assumed to remain the same. The explicit idiosyncratic and implicit conceptual definitions were also coded as to the keywords used. If the definition was implicit through measurement or referential to another researcher, keyword analysis was not done. This enabled the categorization of the amount of alignment between definitions and their operationalization through measures.

Reliability. According to the *Standards for Educational and Psychological Testing* (APA/AERA/NCME 1999, 2014), reliability of data should be established for the current dataset except in the case of a standardized measure. Presence of stability, internal consistency and interrater reliability are a necessary condition for quality studies (Kimberlin and Winterstein 2008). The presence or absence of such documentation was coded for each study. In the case of qualitative studies, a discussion of evidence of rigor, such as triangulation, verbatim participant transcripts, and multiple researcher analysis, was used to indicate the presence of reliability (Bashir et al. 2008). Reliability was coded as (a) present for the current dataset except in the case of a normed test, (b) present from previous research, (c) present from a combination of current and previous datasets, or (d) absent.

Validity. For defensible inferences, valid measures must be used (Crocker and Algina 2006). Cook and Beckman noted that "validity is a property of inferences, not instruments; valildity must be established for each intended interpretation" (p. 166.e8). If measures of content, predictive, concurrent, or construct validity were documented, the validity was coded as *present* versus *absent.* The validity was coded as (a) present for the current dataset, (b) present from previous dataset, (c) present for both current and previous studies, or (d) absent.

Commonality of measure use. The types of measures were categorized as to *commonality*, defined as whether or not the measure was created by the researcher specifically for this study, or was a previously used measure. This is important because using reliable, previously validated measures, instruments, and scales is valued in research (Kimberlin and Winterstein 2008). Each measure was coded as to whether it was study-specific by the researcher, standardized, or a commonly used measure. An instrument that was used in at least one other published study was coded as *commonly used;* a study that used a standardized, normed, commercial instrument was coded as *standardized*; a study that was designed for the study was coded as *study-specific.* An example of a researcher-made study-specific measure was the Cognitive Performance Instrument developed by Johnson et al. (2008) for their study of critical thinking during chemical warfare triage. Researchers McAllister et al. (2009) developed a study-specific evaluation tool with criteria for clinical reasoning in emergency nurses for their study. The California Critical Thinking Skills Test was a standardized test in this coding scheme.

10.3 Results and Discussion

10.3.1 Research Question 1: What Is the Nature of Contextual Variables in Critical Thinking Studies?

In order to evaluate the findings regarding the contextual aspects, the purpose, participants, and research design findings will be reported and discussed.

Purpose. Nearly half (49 %) of the articles in the resulting pool of 43 articles were about critical thinking, and 51 % had clinical reasoning as the focus.

Evaluation or description of a teaching technique was the most common category of purpose studied in the pool, with 63 % of the studies, followed by 26 % of studies done for evaluation or description of students, faculty or providers, 7 % for evaluation of a program of education, and 5 % for evaluation of admission, course performance, or progression decisions.

The purpose of most of these studies was to evaluate teaching strategies. The lack of explicit, nonequivalent, domain-specific definitions and standardized or commonly used domain-specific measures may be limiting the purposes that can be researched.

Participants. The domains of the participants examined by the studies were distributed as follows: 53 % nursing, 26 %, medicine, 5 % Occupational Therapy, Interdisciplinary 5 %, Veterinary 5 %, and Other 5 %. The level of professional experience of the participants varied across the professional life span, ranging from student to resident/new grad to practicing professional. Over half of the studies examined participants at the student level (51 %), 35 % studied participants at the provider level of experience, 12 % examined residents or new grads, and 2 % studied participants at more than one level of experience. The sample sizes ranged from 6 to 2144 participants. Thirty-seven percent of the articles had small sample sizes, 40 % had moderate sample sizes, and 23 % had large sample sizes.

There were several aspects of the participants in the samples that were notable. There were differences in the use of the main constructs between the domains. There is a difference in the amount of usage of CT compared to CR between medicine and nursing domains, in that medicine is mostly focused on CR, whereas nursing research included both terms. Possible explanations might be that nursing research in an earlier developmental stage, or that nursing include more categories of behavior outside diagnosis, such as resource allocation, interpersonal negotiation, stronger patient relationships, as well as clinical knowledge of best practice for the situation. Nursing protests the fact that caring and integrative functions of nursing care are often not included in measures (Tanner 1997), and medicine complains of the focus on diagnostic reasoning and not other aspects of medicine (Norman 2005).

To really understand critical thinking we must look at its initial development and refinement across the professional life span; however, most of the studies focused on students only. This is tied to a sampling problem; very little purposive sampling was done; most studies were conducted at only one site, and researchers were investigating participants at their own institution.

Research Designs. The studies were categorized according to the research design categories described in Creswell (1994, 2014): experimental, quasi-experimental, preexperimental, and qualitative. Experimental studies comprised 16 % of the studies, quasi-experimental studies comprised 28 % of the studies, preexperimental studies comprised 33 % of the studies, and qualitative studies comprised 23 % of the studies.

This finding quantifies a trend noted in previous studies (Johnson 2004) of a very small number of experimental studies. Without experiments it is difficult to test pedagogies to produce an evidence base for teaching.

10.3.2 *Research Question 2: How and How Well Are Definitions for Critical Thinking and Clinical Reasoning Specified in the Literature?*

Clarity: Explicit definitions were used in 58 % of the studies (Table 10.2). Sixty-eight percent of the explicit definitions were shared, and 40 % of all definitions were explicit shared, so this was the largest subcategory of definitions. Thirty-two percent of the explicit definitions were idiosyncratic. For all of the studies using implicit definitions, 39 % used implied-*conceptual* definitions, 33 % used implied-*referential*, and 28 % used implicit through *measurement* definitions.

Interrater reliability was established for the coding of *clarity* of the definitions. Interrater reliability was established using two raters, the author and a second rater who is a doctorally prepared scientist. After discussion of the codebook, a random sample of 10 % of the studies was examined until high interrater agreement was obtained ($\alpha = 0.90$).

As noted in Table 10.2, the results indicated that for clarity of definitions, a very large minority of studies, 42 %, did not explicitly state the definition used in the study. Although this has been mentioned in the literature as a problem (Ratanawongsa et al. 2008), this review documents the great extent to which this is an issue for the critical thinking literature. It also makes clear that there is a greater consensus on explicit definitions of critical thinking (73 % of explicit definitions were shared compared to 60 % for clinical reasoning), but this is largely because standardized instruments were used to a much greater extent in CT studies. In the studies where definitions were implicit, half of the CT studies were referential, using a previous source for the definition, whereas CR implicit studies were 50 % conceptual. It may be that using concepts as opposed to the instrument or an outside source for the definition is more appropriate and better quality strategy to indicate a study's definition of critical thinking, so this trend may be one of several that may indicate greater development in study quality for CR studies.

Table 10.2 Frequency and relative percent of definitions by clarity category and construct

Definitional category	Construct			
	Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)
Explicit	15	–	10	–
Shared	11	73	6	60
Idiosyn	4	27	4	40
Implicit	6	–	12	–
Conceptual	1	17	6	50
Referential	2	33	4	33
Measure	3	50	2	17
Total all	21	100	22	–

Note Idiosyn is the abbreviation for idiosyncratic

Table 10.3 Frequency and percentage of domain specificity of clinical thinking studies by construct

			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Specific	24	56	5	24	19	86
General	19	44	16	76	3	14
Total	43	100	21	100	22	100

Domain specificity. The domain specificity of 44 % of definitions for all studies were domain-general, 56 % were domain-specific. Within the constructs of CR and CT the distribution was very different; 86 % of clinical reasoning studies had domain-specific definitions, whereas 76 % of critical thinking studies were domain-general.

This may be because although CT and CR are often equated, to really measure CR a domain-specific definition is usually adopted as more aspects of patient care as opposed to just cognitive functions are captured. This was seen in the definition data entries; for example, Kuiper et al. (2008) defined CR as "gathering information about many different aspects of the clinical situation, perceiving and interpreting salient cues, and then, on the basis of those cues and relevant knowledge, choosing the treatment strategy that is most likely to facilitate goal achievement (Rogers and Holm 1983)," which includes a lot of patient specific data, whereas a typical CT definition was Drennan's (2010), "identify central issues and assumptions in an argument, recognize important relationships, make correct inferences from data, deduce conclusions from information or data provided, interpret whether conclusions are warranted on the basis of the data given, and evaluate evidence…" (p. 423), which is less patient-centric (Table 10.3).

This analysis of domain specificity of definitions can advance the conversation in measuring critical thinking by ensuring researchers consider this important aspect of the definition they choose to operationalize. This provides measurement of a concept that was a frequent explanation for the lack of consistency in the results for critical thinking. As cited in the introduction, many authors have cited a need for domain-specific instruments.

Equivalency. Most of the studies reported some equivalence between the definitions of critical thinking terms (Table 10.4). For example, Cruz et al. (2009) stated "the hypothesis that a 4-day course consisting of 16 h of content with discussion of critical thinking and clinical reasoning would positively influence participant's diagnostic accuracy scores." Further discussion made it clear that CR was seen as equivalent to CT. Seventy percent reported the presence of equivalency of definitions between at least two of all the terms, and this was even more true for clinical reasoning (77 %) than for critical thinking (62 %) studies.

Although the interchangeable use of critical thinking terms has been frequently cited in the literature, it was surprising that a full 70 % of studies attributed

Table 10.4 Frequency and percentage of equivalency of term usage presence in clinical thinking studies by construct

Equivalency presence			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Present	30	70	13	62	17	77
Absent	13	30	8	38	5	23
Total	43	100	–	–	–	–

some degree of equivalency between critical thinking, clinical reasoning, problem-solving, decision-making, and clinical judgment, especially in light of the findings above about differences in clarity and domain specificity of the definitions. It is notable that the equivalency was noted to an even larger degree in CR studies. Further study is warranted to clarify exactly which terms are being considered equivalent.

10.3.3 Research Question 3: How and How Well Are Measures of Critical Thinking and Clinical Reasoning Specified in the Literature?

Commonality. *Study-specific* measures were used for 40 % of the studies (Table 10.5). *Commonly used* measures were used 37 % of the time. Commonly used measures included Script Concordance Test, Key Feature Exam, and Health Sciences Reasoning Test. *Standardized* measures were used 23 % of the time. These included the California Critical Thinking Skills Test, the ATI Critical Thinking Test, and the Watson-Glaser Critical Thinking Test. For the CT articles, standardized measures were used 48 % of the time, but 0 % of the time for the CR studies. Also, 55 % of the studies used study-specific measures in CR as compared to 24 % for CT.

Table 10.5 Frequency and percentage of commonality of measure usage among clinical thinking studies by construct

Measure commonality			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Study-Specific	17	40	5	24	12	55
Commonly used	16	37	6	29	10	45
Standardized	10	23	10	48	0	0
Total	43	100	21	100	22	100

Table 10.6 Frequency and percentage of alignment of definitions with measures of clinical thinking

Category			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Full	20	47	11	52	9	41
Partial	10	23	4	19	6	27
NA	13	30	6	29	7	32
Total	43	100	21	100	22	100

Researchers are probably creating new measures more than necessary or optimal; perhaps searching for instruments from other domains could lead to a larger stable of domain-specific, well-validated instruments.

Alignment. For the measure of alignment adapted from Alexander and Murphy (2000) and Dinsmore et al. (2008), less than half the studies, 47 %, had a full alignment between the definition of critical thinking and the operationalization through the measure. As Table 10.6 shows, partial alignment occurred in 16 % of the studies, and minimal alignment occurred in 7 % of the studies. Thirty percent of the studies were excluded because the definition was implicit referential or implicit measurement, or unable to be determined because the measurement was not adequately described. Interrater reliability was established for the coding of *alignment* of the measures at $\alpha = 0.90$.

Operationalization of the construct was a big problem for this pool of studies, with less than half having complete alignment. This distribution was about the same for both CR and CT. The lack of clarity in definitions, or lack of explicit definitions, is probably driving this. Researchers often listed a menu of definitions or attributes of critical thinking without specifying which definition or model of thinking they were using to measure critical thinking.

Reliability. Reliability documentation was absent in 7 % of studies (Table 10.7). Reliability was obtained from a previous dataset in 23 % of studies. Reliability was obtained for the current sample in 54 % of studies. Reliability was obtained for both

Table 10.7 Frequency and percentage of reliability categories of studies by construct

Reliability documentation			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Absent	3	7	1	5	2	9
Previous sample only	10	23	7	33	3	14
Current sample only	23	54	8	38	15	68
Both current and previous samples	7	16	5	24	2	9
Total	43	100	21	100	22	100

Table 10.8 Frequency and Percentage of validity of studies by construct

Validity documentation			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Previous sample only	21	49	15	71	6	27
Current sample only	18	42	4	19	14	64
Both current and previous samples	4	9	2	10	2	9
Total	43	100	21	100	22	100

Table 10.9 Frequency and percentage of studies at different levels of analysis

Analysis level			Construct			
	All studies		Critical thinking		Clinical reasoning	
	n	Percent (%)	*n*	Percent (%)	*n*	Percent (%)
Descriptive, narrative	12	28	7	32	5	24
t-test, ANOVA	30	70	15	68	15	71
Regression, factor analysis	1	2	0	0	1	5
Total	43	100	22	100	21	100

the current and a previous sample in 16 % of studies. A larger percentage reported reliability for both the current and previous datasets for CT studies (24 %) compared to CR studies (9 %), again driven by the use of standardized tests with documented reliability.

Validity. Validity was obtained from a previous sample in 49 % of studies (Table 10.8). Validity was obtained for the current sample only in 42 % of studies. And validity was obtained for both the current and a previous sample in 9 % of studies. All studies documented validity in some way. Nearly half based the evaluation of validity on a previous sample. There was a notable difference in the constructs; most of the CR studies used validity from the current sample only, whereas for CT most of the studies used validity from previous research only, again probably driven by the extensive use of normed, standardized tests in CT research. Only 9 % of the studies documented validity from both previous *and* current datasets, which is best practice in research reporting (APA/AERA/NCME 1999, 2014) (Table 10.9).

10.4 Conclusions and Future Directions

There were several limitations to the generalizability of this study. It was conducted by a single researcher, with a second researcher for interrater reliability of key variables. Systematic review technology calls for a team of researchers at each step of the process from electronic searching to conclusions, as well as a technical advisory panel to provide input.

The search terms were limited to nursing and medicine. More professions could be included, and the differences and similarities between the professions could be more closely examined. There is a current professional groundswell to promote more interprofessionalism, and knowledge of interprofessional differences in research approaches to measuring critical thinking and clinical reasoning would be part of a foundation for that. The experience level of practicing professionals not always clearly described in the studies; this could well have an influence on findings if experts are combined with more novice providers. Also, there may be international differences related to scope of practice and cultural differences among countries.

It was not possible to compute effect size due to limitations in reporting: Very few studies reported means, SD, and *n*, needed in order to have higher quality evidence with meta-analyses. Although most studies used inferential statistics such as *t*-test and ANOVA in the analysis, regression analysis and factor analysis to validate scales or regression to test models were almost never used. The lack of sampling in nearly all the studies also decreases the quality and generalizability of results.

The limitation of the search terms to clinical reasoning and critical thinking could have provided a bias in the findings; future systematic reviews of critical thinking should include problem-solving, clinical judgment, and decision-making to help determine the key attributes of each, and move toward less equivalency in the use of the terms. Finally, only one database was used, PsycINFO. Although PsycINFO is a highly respected comprehensive database, it is possible that an emphasis on studies of students resulted due to the journals included. Future systematic reviews of this topic could include MEDLINE and CINAHL databases also. Bias could be further reduced with hand searching that included key journals and work of key researchers.

It is clear that critical thinking and clinical reasoning are still important terms of study for researchers in the health professions. This systematic review of studies of critical thinking in the health professions suggests that definitional conflation and lack of measurement alignment limited the generalizability and value of the study pool findings. Based on this research, some suggestions for future research can be made.

More use of purposive samples spanning multiple levels of experience (student, new grad, very experienced provider) may lead to increased generalizability of findings. Researchers should explicitly state the conceptual and operational definitions used for measures of critical thinking, and be sure that the attributes measured in the instrument line up with the attributes of the definition. Researchers should be aware of the implications of domain-specific or general definitions and consider the impact on outcomes. The role of knowledge may be key in the distinction, as noted in Norman (2005); because many of the strategies to improve CR assess different kinds of knowledge, such a conceptual, procedural, or conditional. If standardized reference definitions for critical thinking and similar terms were developed and disseminated, a subsequent analysis of attributes could allow for distinctions between the attributes of the terms.

Implications for assessments of reliability and validity documentation should include both previous and current datasets as appropriate. If journals that do not currently adhere to the AERA/APA/NCME standards for reliability and validity changed their policies, this could change quickly. Institutional supports for adequate research time, and faculty development in psychometrics could also help increase the methodological quality of research. Kane outlines an argument-based approach to validity, which could be used to ensure better alignment of definitions with measures (Kane 1992). An analysis of keywords to clarify exactly which attributes are being used to define CT or CR would help illuminate which dimensions are being measured. It is also possible that critical thinking components will vary by subdomains such as maternity nursing or surgical medicine, so subdomain analysis would be helpful. Future studies should rate and grade the evidence, not just describe it. Coding schemes such as developed for this study could be used. There could be international differences in use of terms.

In conclusion, articulation of the attributes of conceptual and operational definitions, as well as reforms in research designs and methodologies, may lead to more effective measurement of critical thinking. This could lead to more advanced research designs and more experimental studies. Faculty could target certain aspects of critical thinking in their pedagogy, such as compiling evidence, or making inferences. As discussed elsewhere in this volume, domain-specific measures, precise operationalization of constructs, and curricular alignment should be embedded in professions education. More effective teaching strategies, curricular changes, and ultimately patient care solutions could result. These principles of improved conceptualization and measurement to improve professional competence can also be applied to professions outside of healthcare. For example, one engineering study examined a critical thinking curriculum in the engineering context that used a strong theoretical base and reliable rubrics to evaluate critical thinking in engineering projects (Ralston and Bays 2013). Professionals can then strive to deliver higher quality outcomes in a complex society that demands not rote thinking, but the highest levels of critical thinking possible.

Issues/Questions for Reflection

- How can the measures used in professions education research be designed to align more effectively with definitions of constructs used in the study?
- How can reliable and valid measures be shared more extensively among researchers?
- What will be the benefits of sharing definitions and measures of performance between professions and disciplines?

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*, 1102–1134. doi:10.3102/0034654308326084

Ajjawi, R., & Higgs, J. (2008). Learning to reason: A journey of professional socialization. *Advances in Health Sciences Education, 13*, 133–150. doi:10.1007/s10459-006-9032-4

Alexander, P. A., & Murphy, P. K. (2000). The research base for APA's learner-centered psychological principles. In N. Lambert & B. McCombs (Eds.), *How students learn* (pp. 25–60). Washington, DC: American Psychological Association.

American Association of Colleges of Nursing. (2008). *Essentials of baccalaureate education for professional nursing practice*. Washington, DC: American Association of Colleges of Nursing.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bashir, M., Afzal, M. T., & Azeem, M. (2008). Reliability and validity of qualitative research. *Pakistani Journal of Statistics and Operation Research, IV*(1), 35–45.

Benner, P. E., Hughes, R. G., & Sutphen, M. (2008). Clinical reasoning, decision-making, and action: Thinking critically and clinically. In R. G. Hughes (Ed.), *Patient safety and quality: An evidence-based handbook for nurses. AHRQ Publication No. 08-0043*. Rockville, MD: Agency for Healthcare Research and Quality.

Blondy, L. C. (2011). Measurement and comparison of nursing faculty members' critical thinking skills. *Western Journal of Nursing Research, 33*, 180–195. doi:10.1177/0193945910381596

Brunt, B. A. (2005). Critical thinking in nursing: An integrated review. *Journal of Continuing Education in Nursing, 36*, 60–67.

Carpenter, C. B., & Doig, J. C. (1988). Assessing critical thinking across the curriculum. *New Directions for Teaching and Learning, 34*, 33–46. doi:10.1002/tl.37219883405

Chan, Z. C. Y. (2013). A systematic review of critical thinking in nursing education. *Nurse Education Today, 33*, 236–240.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*, 166.e7–166.e16.

Cooke, M., Irby, D. M., & O'Brien, B. C. (2010). *Educating physicians: A call for reform of medical school and residency*. San Francisco: Jossey-Bass.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.

Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Wadsworth Group/Thomas Learning.

Cruz, D. M., Pimenta, C. M., & Lunney, M. (2009). Improving critical thinking and clinical reasoning with a continuing education course. *The Journal of Continuing Education in Nursing, 40*, 121–127. doi:10.3928/00220124-20090301-05

Dinsmore, D. H., Alexander, P. A., & Loughlin, S. M. (2008). The impact of new learning environments in an engineering design course. *Instructional Science, 36*, 375–393. doi:10.1007/s11251-008-9061-x

Drennan, J. (2010). Critical thinking as an outcome of a master's degree in nursing programme. *Journal of Advanced Nursing, 66*, 422–431. doi:10.1111/j.1365-2648.2009.05170.x

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*, 4–10. doi:10.3102/0013189X018003004

Ennis, R. H. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy, 14*, 5–24.

Evidence for Policy and Practice Information (EPPI). (N.D.). *Different types of review*. Retrieved from http://eppi.ioe.ac.uk/cms/default.aspx?tabid=1915&language=en-US

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: The California Academic Press.

Forneris, S. G., & Peden-McAlpine, C. (2007). Evaluation of a reflective learning intervention to improve critical thinking in novice nurses. *Journal of Advanced Nursing, 57*, 410–421. doi:10.1111/j.1365-2648.2006.04120.x

Funkesson, K. H., Anbäckena, E.-M., & Ek, A.-C. (2007). Nurses' reasoning process during care planning taking pressure ulcer prevention as an example: A think-aloud study. *International Journal of Nursing Studies, 44*, 1109–1119. doi:10.1016/j.ijnurstu.2006.04.016

Göransson, K. E., Ehnfors, M., Fonteyn, M. E., & Ehrenberg, A. (2007). Thinking strategies used by registered nurses during emergency department triage. *Journal of Advanced Nursing, 61*, 163–172. doi:10.1111/j.1365-2648.2007.04473.x

Harden, R. M., Grant, J., Buckley, G., & Hart, I. R. (2000). Best evidence medical education. *Advances in Health Sciences Education, 5*, 71–90. doi:10.1023/A:1009896431203

Higgs, J., & Jones, M. (Eds.). (2000). *Clinical reasoning for health professionals*. Boston: Butterworth Heinemann.

Huang, G. C., Newman, L. R., & Schwartzstein, R. M. (2014). Critical thinking in health professions education: Summary and consensus statements of the millennium conference 2011. *Teaching and Learning in Medicine, 26*(1), 95–102.

Institute of Medicine. (2010). *The future of nursing: Leading change, advancing health*. Washington, D.C.: National Academies Press.

Interprofessional Education Collaborative Expert Panel. (2011). *Core competencies for interprofessional collaborative practice: Report of an expert panel*. Washington, D.C.: Interprofessional Education Collaborative.

Johansson, M. E., Pilhammar, E., & Willman, A. (2009). Nurses' clinical reasoning concerning management of peripheral venous cannulae. *Journal of Clinical Nursing, 18*, 3366–3375. doi:10.1111/j.1365-2702.2009.02973.x

Johnson, M. (2004). What's wrong with nursing education research? *Nurse Education Today, 24*, 585–588. doi:10.1016/j.nedt.2004.09.005

Johnson, D., Flagg, A., & Dremsa, T. L. (2008). Effects of using human patient simulator (HPS™) versus a CD-ROM on cognition and critical thinking. *Medical Education Online, 13*(1), 1–9. doi:10.3885/meo.2008.T0000118

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. doi:10.1037/0033-2909.112.3.527

Kataoka-Yahiro, M., & Saylor, C. (1994). A critical thinking model for nursing judgment. *Journal of Nursing Education, 33*, 351–356.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*, 2276–2284.

Krupat, E., Sprague, J. M., Wolpaw, D., Haidet, P., Hatem, D., & O'Brien, B. (2011). Thinking critically about critical thinking: Ability, disposition or both? *Medical Education, 45*(6), 625–635.

Kuiper, R. A., Heinrich, C., Matthias, A., Graham, M. J., & Bell-Kotwall, L. (2008). Debriefing with the OPT model of clinical reasoning during high fidelity patient simulation. *International Journal of Nursing Education Scholarship, 5*(1), Article 17. doi: 10.2202/1548-923X.1466

Mamede, S., Schmidt, H. G., Rikers, R. M., Penaforte, J. C., & Coelho-Filho, J. M. (2007). Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical Education, 41*, 1185–1192. doi:10.1111/j.1365-2923.2007.02921.x

McAllister, M., Billett, S., Moyle, W., & Zimmer-Gembeck, M. (2009). Use of a think-aloud procedure to explore the relationship between clinical reasoning and solution-focused training in self-harm for emergency nurses. *Journal of Psychiatric and Mental Health Nursing, 16*, 121–128. doi:10.1111/j.1365-2850.2008.01339.x

McCartney, K., Burchinal, M. R., & Bub, K. L. (2006). Best practices in quantitative methods for developmentalists. *Monographs of the Society for Research in Child Development, 71,* 285. Chapter II, Measurement issues. doi: 10.1111/j.1540-5834.2006.00403.x

Nikopoulou-Smyrni, P., & Nikopoulos, C. K. (2007). A new integrated model of clinical reasoning: Development, description and preliminary assessment in patients with stroke. *Disability and Rehabilitation, 29*, 1129–1138. doi:10.1080/09638280600948318

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*, 418–427. doi:10.1111/j.1365-2929.2005.02127.x

Ovretviet, J. (2011). BMJ Qual Saf 2011. *20*(Suppl1), i18ei23. doi:10.1136/bmjqs.2010.045955

Patel, V. L., Arocha, J. F., & Zhang, J. (2004). Thinking and reasoning in medicine. In Keith Holyoak (Ed.), *Cambridge handbook of thinking and reasoning*. Cambridge, UK: Cambridge University Press.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor, MI: University of Michigan National Center for Research to Improve Postsecondary Teaching and Learning.

Ralston, P. A., & Bays, C. L. (2013). Enhancing critical thinking across the undergraduate experience: an exemplar from engineering. *American Journal of Engineering Education (AJEE), 4*(2), 119–126.

Ratanawongsa, N., Thomas, P. A., Marinopoulos, S. S., Dorman, T., Wilson, L. M., Ashar, B. H., et al. (2008). The reported reliability and validity of methods for evaluating continuing medical education: A systematic review. *Academic Medicine, 83*, 274–283. doi:10.1097/ACM.0b013e3181637925

Rogers, J. C., & Holm, M. B. (1983). Clinical reasoning: The ethics, science, and art. *American Journal of Occupational Therapy, 37*, 601–616. doi:10.5014/ajot.37.9.601

Ross, D., Loeffler, K., Schipper, S., Vandermeer, B., & Allan, G. M. (2013). Do scores on three commonly used measures of critical thinking correlate with academic success of health professions trainees? A systematic review and meta-analysis. *Academic Medicine, 88*(5), 724–734.

Schell, R., & Kaufman, D. (2009). Critical thinking in a collaborative online PBL tutorial. *Journal of Educational Computing Research, 41*, 155–170. doi:10.2190/EC.41.2.b

Simmons, B. (2010). Clinical reasoning: Concept analysis. *Journal of Advanced Nursing, 66*, 1151–1158. doi:10.1111/j.1365-2648.2010.05262.x

Simpson, E., & Courtney, M. D. (2002). Critical thinking in nursing education: A literature review. *International Journal of Nursing Practice, 8*(April), 89–98. doi:10.1046/j.1440-172x.2002.00340.x

Tanner, C. A. (1997). Spock would have been a terrible nurse and other issues related to critical thinking. *Journal of Nursing Education, 36*, 3–4.

Turner, P. (2005). Critical thinking in nursing education and practice as defined in the literature. *Nursing Education Perspectives, 26*, 272–277.

Walsh, C. M., & Seldomridge, L. A. (2006a). Critical thinking: Back to square two. *Journal of Nursing Education, 45*, 212–219.

Walsh, C. M., & Seldomridge, L. A. (2006b). Measuring critical thinking: One step forward, one step back. *Nurse Educator, 31*, 159–162.

Waltz, C. F. (2010). *Measurement in nursing and health research* (4th ed.). Philadelphia: Springer.

Watson, R., Stimpson, A., Topping, A., & Porock, D. (2002). Clinical competence assessment in nursing: A systematic review. *Journal of Advanced Nursing, 39*, 421–431.

Wolpaw, T., Papp, K. K., & Bordage, G. (2009). Using SNAPPS to facilitate the expression of clinical reasoning and uncertainties: A randomized comparison group trial. *Academic Medicine, 84*, 517–524. doi:10.1097/ACM.0b013e31819a8cbf

Context (n. d.). In Merriam-Webster.com. Retrieved April 11, 2012 from http://www.merriam-webster.com/dictionary/context.

Alexander, P. A., Dinsmore, D. L., Fox, E., Grossnickle, E., Loughlin, S. M., Maggioni, L., Parkinson, M. M., & Winters, F. I. (2011). Higher order thinking and knowledge: Domain-general and domain-specific trends and future directions. In G. Shraw & D. H. Robinson (Eds.), Assessment of higher order thinking skills (pp. 47-88). Charlotte, NC: Information Age Publishing.

Fonteyn, M. E., & Cahill, M. (1998). The use of clinical logs to improve nursing students' metacognition: A pilot study. Journal of Advanced Nursing, 28(1), 149-154.

Chapter 11
Understanding the Assessment of Clinical Reasoning

Joseph Rencic, Steven J. Durning, Eric Holmboe and Larry D. Gruppen

Abstract Clinical reasoning assessment is an essential component of determining a health professional's competence. Clinical reasoning cannot be assessed directly. It must be gleaned from a health professional's choices and decisions. Clinical knowledge and knowledge organization, rather than a general problem solving process, serve as the substrate for clinical reasoning ability. Unfortunately, the lack of a gold standard for the clinical reasoning process and the observation of context specificity make it difficult to assess clinical reasoning. Information processing theory, which focuses on the way the brain processes and organizes knowledge, has provided valuable insights into the cognitive psychology of diagnostic and therapeutic reasoning but failed to explain the variance in health professional's diagnostic performance. Situativity theory has emerged suggesting that this variance relates to context-specific factors that impact a health professional's clinical reasoning performance. Both information processing and situativity theory inform the way in which we assess clinical reasoning. Current assessment methods focus on standardized testing of knowledge to maximize psychometric parameters and work-based assessments which evaluate clinical reasoning under authentic, uncertain conditions that can decrease the reliability of these measurements. Issues of inter-rater reliability and context specificity require that multiple raters assess multiple encounters in multiple contexts to optimize validity and reliability. No single assessment method can assess all aspects of clinical reasoning; therefore, in order to improve the quality of assessments of clinical reasoning ability, different combinations of methods that measure different components of the clinical reasoning process are needed.

J. Rencic (✉)
Tufts Medical Center, Boston, USA
e-mail: jrencic@tuftsmedicalcenter.org

S.J. Durning
Uniformed Services University, Bethesda, USA

E. Holmboe
Accreditation Council of Graduate Medical Education, Chicago, Illinois, USA

L.D. Gruppen
University of Michigan Medical School, Ann Arbor, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_11

Takeaway

- Clinical reasoning ability depends on a health professional's knowledge and knowledge organization rather than a general thinking process.
- Clinical reasoning is context-specific; clinician or trainee characteristics account for only a small amount of the variance in diagnostic accuracy.
- Determining the validity of any clinical reasoning assessment method is challenging due in part to the situation specific nature of clinical reasoning (context specificity).
- Most clinical reasoning assessment methods provide adequate reliability for a high stakes examinations provided adequate sampling.
- No gold standard for the clinical reasoning process exists.
- High-stakes assessment focus on the accuracy of diagnostic and therapeutic choices because "process" checklists can downgrade advanced trainees or experienced health professionals.
- There is no "magic bullet" for assessing clinical reasoning; therefore, multiple assessments by multiple raters in diverse contexts is key to adequate clinical reasoning assessment.
- We believe that a combination of knowledge assessment (e.g., multiple choice question examination) and clinical skills assessment (e.g., 10 station objective structured clinical examination) should be used in high-stakes conditions.

Clinical reasoning has been viewed by some as the "Holy Grail" of assessment (Schuwirth 2009). It lies at the foundation of clinical performance and therefore should be a key focus of any assessment framework for health professionals. A number of significant challenges exist in assessing clinical reasoning. Consider the following example.

> Imagine you are a physician on the general internal medicine service working with a fourth year student on his acting internship (a month-long practice rotation for being a "real" intern). You become concerned because the student's patient assessments suggest an inability to synthesize the data into a coherent whole, to create a prioritized differential diagnosis, and to select the correct diagnosis. You believe that he can't see the forest for the trees and lacks an ability to match the patient's problems with the correct diagnosis. He's done nothing egregiously wrong but you do not think he is capable of becoming an intern without significant remediation. You contact the medical school and discover his end of rotation exam and USMLE step 1 scores were all greater than the 90th percentile nationally. What assessment tools can you use to gain further insight into the student's deficits? On your end-of-rotation summative evaluation, what "proof" of poor clinical reasoning can you provide to demonstrate that he should fail the rotation? Assuming the pattern continues, what evidence of poor clinical reasoning will the medical school need to produce to justify his dismissal?

In this chapter, we will attempt to provide some answers to the questions posed in the example above. We start the chapter with a working definition of clinical reasoning because it is difficult to assess something without knowing what it is and is not. Next we consider the history of clinical reasoning assessment because it provides a perspective that will help to understand current assessment strategies. We then review the components of diagnostic reasoning and how these are employed during a clinical encounter. This section has practical implications on how a teacher might remediate the student described above. Finally, we highlight key issues specific to clinical reasoning and describe tools currently used for its assessment. We conclude that no perfect tool to assess clinical reasoning exists but by combining multiple tools over many observations, researchers and teachers can obtain a more complete approximation of clinical reasoning ability.

11.1 Defining Clinical Reasoning

Discussing something as complex as clinical reasoning assessment must start with definitions and distinctions. Defining "clinical reasoning" is a major challenge because it means different things to different researchers, practitioners, and teachers in different health professions. There are numerous terms that are subsumed by "clinical reasoning," serve as synonyms, or denote related but conceptually distinct phenomena. The definitional challenge is increased by the diverse intellectual origins of the research methods and background literature and theories that are used to understand "clinical reasoning." Attaining consensus on a canonical definition for clinical reasoning is not the goal of this chapter. Rather, we will focus on a more inclusive framework that can be summarized as *the cognitive and physical processes by which a health care professional consciously and unconsciously interacts with the patient and environment to collect and interpret patient data, weigh the benefits and risks of actions, and understand patient preferences to determine a working diagnostic and therapeutic management plan whose purpose is to improve a patient's well-being.*

This working definition incorporates both diagnostic reasoning and therapeutic reasoning, although these two constructs have each been characterized by different research methods and literature. Diagnostic reasoning (i.e., the thinking process that seeks to classify a patient case into one of several potential diagnostic categories and ends when that decision has been made) has typically been described by psychological theories of problem solving (Newell and Simon 1972), expert performance (Ericsson et al. 2006), and knowledge representation (Bordage 2007; Schmidt and Rikers 2007). It is often studied with the methods developed in cognitive psychology to assess knowledge acquisition, knowledge organization, problem representation, and problem solving. Therapeutic reasoning (i.e., the thinking process that leads to and includes a treatment decision) has been characterized primarily, although not exclusively, through decision analytic frameworks, which assess how professionals make decisions when "… a set of fixed alternatives are contrasted with a normative model based on probability theory, indicating optimal choices under conditions of uncertainty" (Johnson et al. 1991). By comparing decisions made by individual participants against those indicated by normative models based on decision trees, research has discovered numerous cognitive biases that impact judgments. (Croskerry 2003; Graber et al. 2005; Tversky and Kahneman 1974).

Despite these distinct traditions, we believe that clinical reasoning should house both diagnostic and therapeutic reasoning while recognizing the potential value of the subdivision from a research perspective. We acknowledge that these two forms of reasoning may be different but related constructs. For example, successful therapeutic reasoning is often dependent upon successful diagnostic reasoning. Future research should link these two constructs more closely, focusing on their similarities rather than their differences and clarifying the shared knowledge and cognitive processes that demands their joint inclusion under the umbrella of clinical reasoning.

11.1.1 Profession-Specific Definitional Considerations

Although all health professions deal in diagnostic and therapeutic realms, the goals of their reasoning approaches vary based on their primary tasks. For example, Fonteyn and Grobe (1993) suggest that nurses focus more on assessing the significance and relevance of patient data to achieving the overall treatment plan for each patient rather than diagnosis per se. The physical and occupational therapy literature describes three types of reasoning: procedural, interactive, and conditional reasoning. These reasoning types serve as "lenses" through which physical therapists' interactions with their patients' can be understood. Procedural reasoning focuses primarily on the disease with the goal of maximizing function. Interactive reasoning focuses on understanding clients' as people and understanding their feelings about therapy. Finally, conditional reasoning applies a holistic approach trying to integrate procedural and interactive reasoning into clients' broader social context and maximize their quality of life based on this context (Liu et al. 2000). Likewise, physical therapists use narrative reasoning (i.e., listening and understanding patients' stories of their illnesses), a component of diagnostic reasoning, to understand patients' conception of their disability and personal preferences (Jones et al. 2008).

Now that clinical reasoning has been defined, we will review the history of clinical reasoning research. This historical perspective will enhance readers' understanding of current conceptions of clinical reasoning assessment.

11.2 Historical and Contemporary Understanding of the Cognitive Processes of Clinical Reasoning

Over the past several decades, a number of approaches to clinical reasoning assessment have been explored in parallel with prevailing theories of the day. Here, we will briefly outline "highlights" of these approaches and the underlying theoretical perspectives that shaped assessment in each period leading to our proposal for a review.

11.2.1 The General Problem Solving Era

In the 1960s, medical education scholars were convinced that clinical reasoning was based on general problem solving skills, which would be superior in experts than in novices. These problem solving skills were thought to be both generalizable and teachable and, if acquired, would result in superior problem solving performance. These predictions were shown to be incorrect. For example, studies using Patient Management Problems (McCarthy and Gonella 1967; Rimoldi 1963) and other long case formats indicated that an expert clinician (who presumably had these generalizable problem solving skills) performed well on one case but did quite poorly on another. Indeed, the correlation of performance between such cases ranged from 0.1 to 0.3 (as a reference, the goal for a reliable test is at least 0.6–0.8) (Elstein et al. 1978). Performance was discovered to be highly dependent on knowledge content within a given domain (Eva 2003; Eva et al. 1998), thus establishing the concept of *content specificit*y. This discovery led to the next era in clinical reasoning theory.

11.2.2 Expertise as Knowledge Organization: The Era of Information Processing

With the demise of general problem solving skills as a viable explanation of clinical reasoning performance, investigators turned in a new direction. Stimulated by the emergence of the computer and artificial intelligence, investigators adopted an information processing model for clinical reasoning (Pauker et al. 1976). Information processing theory focused attention on knowledge organization in memory, rather than generalized problem solving processes or algorithms (Elstein et al. 1978). Investigators explored numerous forms of knowledge representation (Rosch 1978) and organization, such as illness scripts (i.e., mental representations of key epidemiological and clinical features that occur with a given diagnosis). Illness scripts were thought to derive from exemplars (i.e., individual patients with the given disease) and prototypes (i.e., a compilation of various key features of a specific disease based on experiences and reading into a general mental representation) (Abbott et al. 1985; Charlin et al. 2007).

Two key theories emerged from information processing: dual process theory and cognitive load theory. Dual process theory describes two distinct approaches to reasoning: nonanalytic and analytic (Kahneman 2011). Nonanalytic reasoning is typically fast, subconscious, and seemingly effortless. It reflects the use of patterns and/or heuristics ("mental rules of thumb") encoded in long-term working memory. In health professions education, we often speak of illness scripts, which contain the features that constitute a diagnosis, the range that is possible for each symptom and clinical finding as well as the most likely value for each symptom and finding. On the other hand, analytic reasoning is described as a slow, conscious, mentally taxing process that involves actively comparing and contrasting alternatives (e.g., diagnoses or treatment options). Hypothetico-deductive reasoning and Bayesian methods are two examples of analytic strategies. Kahneman (2011) explains that the default pathway for reasoning is nonanalytic, with analytic reasoning being recruited primarily for complicated or unfamiliar problems or situations. He explored heuristics and biases extensively in his research, illuminating their nonrational and nonprobabilistic nature (Tversky and Kahneman 1974). Indeed, a considerable body of research focused on describing and understanding various heuristics and biases during this information processing era.

By studying constraints on the brain's processing capacity, a related information processing theory emerged called Cognitive Load Theory. Cognitive Load Theory refers to the limitations of our human cognitive architecture (Sweller et al. 1998). The brain can only process a limited amount of information at a given time (7 ± 2 items) (Miller 1956). Thus, the mind manages cognitive load through chunking knowledge and/or experiences to liberate cognitive space to focus on the most critical aspects of a given situation. Interestingly, human beings seem to have an enormous capacity for each individual chunk (through long-term memory) and despite the size of a given chunk, the amount of information still occupies one of the 4–7 "slots" or spaces of working memory. A computer analogy may help illustrate this principal. Working memory (the processor) can only actively analyze 4–7 pieces of information at a time (e.g., word documents) but the length of each word document can be tiny or enormous and still take up only one of the slots.

Information processing theory became the preeminent conceptual framework for researchers of clinical reasoning, but this approach encountered its own limits. Chief among these was its inability to explain the variance in performance by an individual clinician on similar cases requiring the same medical knowledge. That is, if a physician had enough medical knowledge to make a diagnosis on one case, why would she fail on a second highly related case? One important answer to this dilemma of "content specificity" lies in researchers neglect of the physical and social environment's impact on a health care provider's clinical reasoning (Elstein et al. 1978); this work led to the era of context specificity.

11.2.3 Clinical Reasoning Expertise as a State: Context Specificity and Situativity Theory

With the difficulties stemming from content specificity, work has recently explored the issue of expertise as a trait (expertise seen as a general or transferable “skill” within a field) as opposed to a state (limited to specific patient circumstances) (Eva et al. 2007). In other words, expertise as a state argues that clinical reasoning performance is specific to the patient, other health professionals on the team, the environmental and their emergent interactions (i.e., the specific situation). Thus, content specificity that was first described by Elstein et al. (1978) has been renamed *context specificity* to capture the notion that something besides the clinical content of a case is influencing diagnoses and therapy.

Situativity theory has emerged to explain context specificity and expand understanding of clinical reasoning beyond the limits of information processing theory (Durning and Artino 2011). Situativity theory posits that the knowledge of the health professional is only one of several, rather than the sole factor that predicts clinical reasoning success. Research shows that clinician factors account for only a small portion of the variance in predicting diagnostic success (Durning et al. 2013). In seeking to explain the rest of this variance, researchers recognized how situativity theory could serve as an explanatory model. Situativity theory stresses the importance of interactions between health care professionals, patients, and environments as fundamental in clinical reasoning (Durning and Artino 2011). It includes situated cognition, distributed cognition, and situated learning. Situated cognition emphasizes the interactions between participants in a situation as well as the environment (Brown et al. 1989). Distributed cognition stresses that deciding and acting in groups is often based on the interaction of thinking with many individuals (e.g., nurses, physical therapists, case managers, and doctors) (Salomon 1993). Situated learning stresses that learning is a result of participation in a community (a community of practice, again focusing on groups) (Lave 1988). Situativity theory suggests that authentic assessments, multifaceted measures (taking into account these different factors) and qualitative assessments are needed (Durning and Artino 2011). One should sample broadly (i.e., many varied situations) to assess clinical reasoning expertise. Validity and reliability are more difficult to establish due to the number of potential factors (e.g., environmental, patient-related) impacting a given clinical situation.

11.3 The Construct and Process of Clinical Reasoning

We endorse a nonlinear view of clinical reasoning and its assessment, given its complex nature. As described previously, a small perturbation in a clinical reasoning assessment exercise (e.g., changing the case presentation while leaving the diagnosis the same) effects diagnostic accuracy (Elstein et al. 1978) much more dramatically than expected in a linear system, suggesting that complex interactions are occurring between the physician, the patient, and environmental factors. A nonlinear view of clinical reasoning espouses multiple correct pathways within boundaries, rather than one "single best pathway" (Durning et al. 2013). We do not mean to imply that we are "rejecting" prior models of clinical reasoning; we believe that there are times that reasoning is linear and times that it is nonlinear. The quality of clinical reasoning thus depends on the ability to flexibly choose from a rich array of strategies and change the strategy(ies) based on the reaction of the patient or on other features in the environment (specifics of the situation or "state dependent").

Thus far, even the most modern methods designed to capture the process of clinical reasoning have not been capable of looking beyond behavior (Schuwirth 2009). For example, think-aloud studies (e.g., cognitive task analysis) are among the common measures applied to date (Elstein et al. 1978; Klein 1998), but they have subjects verbalize their thought processes, which constitutes an indirect, quite possibly reactive, and biased account. Think-aloud protocols are not direct observations of the brain activity that characterize thought processes (Schuwirth 2009), but may interfere with the very mental processes they are intended to assess. Thus, assessment strategies that enable more direct exploration of cognitive processes (e.g., functional MRI), without interfering with these processes, offer advantages to our understanding of reasoning.

The complex interactions that occur during clinical reasoning make assessment a daunting task. However, from a practical standpoint, some means of breaking down the process are essential. We have already discussed the somewhat artificial division of clinical reasoning into diagnostic and therapeutic reasoning. Within these subdivisions, the clinical reasoning process has been further subdivided. For diagnostic reasoning, the literature emphasizes the following steps: data collection, problem representation, hypothesis generation, hypothesis refinement, and working diagnosis/illness script selection. For therapeutic reasoning, the components include probabilities of possible outcomes of treatment or no treatment, the utility or value of those outcomes (e.g., death = 0), relevant patient characteristics, and patient preferences. The "correct answer" (i.e., treatment) for common illnesses with adequate research studies can be determined for a given clinical scenario if adequate data, software, and time for a literature review exist. In these circumstances, assessment is straightforward. Unfortunately, many, if not most, treatment decisions are so complex and context specific that evidence is lacking for a definitively "correct" answer. We will not discuss assessment of therapeutic reasoning further given the limited research in the cognitive psychology of therapeutic reasoning. We will ignore the critical influence of patient and environment in the process for the moment and focus primarily on the clinician's cognitive processes to improve clarity, although this approach represents a simplification.

In its most basic sense, diagnostic reasoning is categorization, a cognitive matching exercise. Diagnosticians apply both nonanalytic (e.g., pattern recognition) and analytic reasoning to categorize a patient's problems into a disease category. We present this process in a linear manner but in reality, we believe it is nonlinear, with a health professional's thought process moving forward and backward iteratively through these components (Gruppen and Frohna 2002). The first component of this process is **data collection**. A health professional obtains a patient's history, performs a physical examination, and orders studies (when necessary) to determine the exact nature of the patient's problem(s). Data collection requires both declarative knowledge ("knowledge **about** things") and procedural knowledge ("knowledge about **how to do** things"). For example, to correctly "collect" aortic stenosis murmur data, a learner must know what aortic stenosis sounds like (declarative knowledge) and how to use the stethoscope to hear it (procedural knowledge). Data collection may be deficient if the learner lacks either knowledge type.

Collected data progressively coalesces into a **problem representation**. A diagnostician attempts to coherently interpret the collected data into a disease category. In diagnostic reasoning, this task can be incredibly challenging. With potentially 20–30 items of patient data to process, distinguishing salient data from "red herrings" requires significant content knowledge and organization. In addition, data that changes the problem representation may emerge over time requiring its "reframing" (Durning and Artino 2011; van Merrienböer and Sweller 2005).

As the problem representation emerges, **hypothesis generation**, or the construction of a differential diagnosis composed of multiple disease categories, occurs. Often hypotheses are triggered by bits of clinical data through nonanalytic reasoning when clinicians recognize a pattern of symptoms in a patient that matches a mental disease representation (i.e., "illness script") within long-term memory. Another term for this triggering effect is "script activation." In complex or unfamiliar patient presentations that lack a clear pattern, clinicians apply analytic reasoning (e.g., differential diagnosis mnemonics like VINDICATE (e.g., **v**ascular conditions, **i**nfections, **n**eoplasm, **d**rugs, **i**mmunologic (autoimmune conditions), **c**ongenital, **a**llergy, **t**rauma, **e**ndocrine/metabolic conditions), pathophysiologic/causal reasoning, point of care resource searches) to enhance hypothesis generation.

Hypothesis generation feeds back to data collection and clinicians gather additional data to discriminate between hypotheses (**hypothesis refinement**). This data collection may include additional history, physical examination, laboratory, or radiological studies. In addition, a diagnostician may use point of care resources to look up information that would aid in the reasoning process (e.g., the typical presentation of the disease, tests of choice, etc.) This portion of the clinical reasoning process conforms to the hypothetico-deductive reasoning model in a nonexperimental sense. That is, given the most likely hypothesis, what symptoms/signs/laboratory/radiological findings are expected and not expected? Of these, which are present or absent in the patient? The reasoning exercise of diagnosis seeks to select and confirm that a health professional's mental construct of the leading disease hypothesis matches the patient's problem representation (i.e., **script selection and confirmation**). If the clinical findings do not match, then the hypothesis is falsified and the process repeats until enough diagnostic certainty is achieved that treatment can begin. At this point the health professional has defined a **working diagnosis**.

11.4 Foundational Issues in Clinical Reasoning Assessment

In Miller's (1990) famous and oft-cited assessment pyramid, knowledge and clinical reasoning figure prominently. Assessments of "knows and "knows how" form the base of the pyramid and mostly target the learner's possession and use of knowledge in standardized, simple uncomplicated contexts (e.g., a testing center). The upper two levels, "shows how" and "does" more specifically target the learner's application of knowledge across multiple situations, integrating clinical reasoning as part of providing care in either a simulated environment or to actual patients.

Assessment of clinical reasoning over the past century has matured substantially in parallel with advancements in the field of psychometrics. Psychometrics enabled the rigorous and efficient development and delivery of high-stakes testing that can robustly target assessment at the "knows, knows how and shows how" levels of the Miller pyramid. However, in recent years there has arisen some backlash against the intense focus on high-stakes testing, even in the health professions (Hodges 2013; Schuwirth and van der Vleuten 2006). One factor in the criticism has been a perceived over-reliance on high-stakes assessment of clinical reasoning at the expense of formative assessment and better feedback as well as other important competencies.

11.4.1 Assessment of Versus for Learning

There is little debate with the aphorism "assessment drives learning" and assessments signal very clearly what programs and educators believe is important (Norcini et al. 2011). The evidence is robust in demonstrating the importance of feedback on professional development (Boud and Molloy 2013; Ericsson 2007; Hattie and Timperley 2007). Assessments of learning can produce some short-term and long-term "educational effect" (Karpicke and Blunt 2011; Larsen et al. 2008). At the undergraduate and graduate levels, testing can focus on higher level questions requiring problem solving and decision-making (Krathwohl 2002), as well as OSCEs stations that assess clinical reasoning. In addition, a number of continuous professional development programs are increasingly incorporating assessment for learning to enhance clinical reasoning (Green et al. 2009; ACP Smart Medicine, n.d.).

11.4.2 Criteria for Good Assessment

Good assessment has the following characteristics: validity or coherence, reproducibility or consistency, equivalence, feasibility, educational effect, catalytic effect (Norcini et al. 2011; van der Vleuten 1996). These criteria should apply to any assessment of clinical reasoning, and each criterion will be weighted differently depending on the purpose of the assessment. For example, a high-stakes, summative licensing exam would need to demonstrate high levels of validity, reproducibility, and equivalence. Formative assessments would logically want to place greater weight on the catalytic effect (i.e., assessment that drives future learning forward) while ensuring a reasonable level of validity and reproducibility. It should be clear that any assessment of clinical reasoning should be "fit for purpose" to maximize its utility and impact, and the importance of purpose should therefore guide what should be measured in regards to clinical reasoning.

11.4.3 What Should We Measure?

When the primary purpose of assessment is to ensure healthcare professionals possess a minimal level of competence for promotion decisions and public accountability, the emphasis has been on measuring the end product of the clinical reasoning process. The best examples are the various licensing and certification examinations given around the globe that test for the "best answer." Typically systematic processes are used to create thresholds that enable discrimination of levels of performance among the test takers; discrimination among trainees is a key purpose of these high-stakes assessments. Assessments that focus on the end product sample from a blueprint representative of the discipline using multiple questions or items. However, there has been increasing use of these types of exams for formative purposes such as in-training examinations in medical residency and fellowship programs and practice tests in medical schools (Babbott et al. 2007; Grabovsky et al. 2014; Chang et al. 2014; National Board of Medical Examiners International Foundations of Medicine, n.d.). The amount and nature of feedback that can be provided with assessments that focus on the end product of clinical reasoning is somewhat limited because the assessments are effectively agnostic to the cognitive process used to generate the response. Process measurement provides the possibility of formative feedback about the reasoning *process*, whereas assessing the *product* of reasoning primarily allows for summative feedback.

11.4.4 How Can We Measure?

Currently our standard methods do not allow us to “see” clinical reasoning, but rather we must infer its presence and character through behaviors. Such inferences require robust sampling over the appropriate domains of knowledge and situations within a discipline. For assessments not conducted in controlled settings and for formats in which context is provided by the assessment task (e.g., written exams and standardized oral exams), the impact of context on the reasoning process cannot be ignored and must be part of the assessment process. Finally, we cannot ignore the nature of the measurement instrument. When the assessment involves a rater who questions and observes it is important to remember the rater *is* the measurement instrument. We know raters are impacted by a host of factors such as their frame of reference, idiosyncrasies, own strengths, and weaknesses in clinical reasoning, bias and local environment (Gingerich et al. 2011; Govaerts et al. 2011; Kogan et al. 2009).

While beyond the scope of this chapter, physiologic measurements are beginning to help us better understand how the neuro-circuitry functions in clinical reasoning in medicine. For example, Durning et al. (2014), using functional MRI imagining, have identified areas of the brain are active during the clinical reasoning process, raising hopes that we may indeed soon have the capacity to actually “see” what is happening in clinical reasoning physiologically.

11.5 Tools for Assessing Clinical Reasoning

We will now focus on specific tools that have been used to assess clinical reasoning. Many of these have been described in the medical literature but are used throughout the health professions. Although the literature suggests a range of reliability for these tools, van Der Vleuten and Schuwirth (2005) have suggested that the key to reliability is adequate time and sampling, rather than a specific tool. Deciding what component(s) of clinical reasoning one wants to measure is the key issue in tool selection. For example, if data collection is a key element of the desired assessment, then an objective structured clinical examination would be more valuable than a multiple choice test. The following sections describe a diverse sampling of clinical reasoning assessment methods.

11.5.1 Standardized Assessment Methods

High-stakes learner assessments require standardization to insure reliable measurements. As a result, licensing bodies, residencies, and medical schools have made significant efforts to standardize learner assessments using psychometrically justifiable tools. We will first focus on such tools that can be used for clinical reasoning assessment.

11.5.1.1 Multiple Choice and Extended Matching Questions

The multiple choice question (MCQ), particularly the single-best answer format (Case and Swanson 2002), is the predominant form of assessment for clinical reasoning in North America across health professions. MCQs are also used in the United Kingdom as part of the qualifications for the Member of the Royal Colleges examinations. Some schools are using the International Foundations of Medicine Examination as part of undergraduate medical education (National Board of Medical Examiners International Foundations of Medicine, n.d.). Finally, the European Union of Medical Specialties is beginning to offer medical knowledge specialty exams in an effort to enhance physician mobility within the European Union.

MCQ questions typically embed clinical context into the question through use of a clinical stem, or vignette, that includes a description of a patient, key clinical features, and "distractors" (i.e., elements irrelevant to the case) (Case and Swanson 2002; Hawkins et al. 2013; Holmboe et al. 2008). Vignette-based MCQs are particularly effective and efficient in assessing the ability of candidates to synthesize multiple pieces of key information and ignore nonessential elements in choosing a diagnosis or treatment.

Extended matching questions (EMQ), developed by the NBME in the 1980s are defined by Case and Swanson (1993) as "*any matching format with more than five options with items are grouped into sets, with a single option list used for all items in a set. A well-constructed extended matching set includes four elements: a theme; a lead-in statement; an option list; and two or more item stems*." (For an example see Case and Swanson 2002, p. 82). Extended matching items are purported to help reduce the cueing effects commonly seen with single-best answer MCQs. These types of questions have been used on the NBME licensing exams and have been shown to possess good psychometric characteristics. However, previous work found that extended matching questions, used on the initial American Board of Internal Medicine examination, did not provide any additional information about ability over a well designed, vignette-based MCQ (Norcini et al. 1982, 1984).

11.5.1.2 Structured (Standardized) Oral Examination

In traditional oral examinations, the examinee performs a history and physical examination of a patient, with or without supplemental laboratory or radiological material, followed by a series of questions from an examiner. The examiner can judge the quality of the history and physical examination (either through a chart or direct observation), and the clinical reasoning process (e.g., asking the examinee to generate a differential diagnosis with supporting evidence) and management plan. It is not “good enough” to simply get the answer “right.” For a number of reasons, such as reliability concerns, rater bias (hawks and doves), standardization of cases, patient fatigue, and logistics, oral examinations with live patients have fallen out of favor in most disciplines. Disciplines that continue to employ high-stakes oral examinations now use more standardized cases (i.e., scripted questions) as the catalyst for oral examinations. One study examined a structured oral examination (SOE) for trainees in neonatology and perinatology and found reasonable reliability for the medical expert role of CanMEDS (Jefferies et al. 2011). Another good example is the American Board of Emergency Medicine’s oral examination for initial certification (Counselman et al. 2014). In this situation the SOE is a high-stakes assessment (Jefferies et al. 2011).

11.5.1.3 Key Features

The key features (KF) approach is designed to specifically target a learner’s decision-making skills with less emphasis on factual recall. KF focuses on how knowledge is used to create a differential diagnosis, decide on what tests or procedures to pursue, choose a management approach, and so on. Bordage and Page (2012) define a key feature as follows:

1. *a critical or essential step(s) in the resolution of a problem,*
2. *a step(s) in which examinees […] are most likely to make errors in the resolution of the problem, or*
3. *a difficult or challenging aspect in the identification and management of the problem in practice.*
   (From the MCC Guidelines for the development of key features problems and test cases. April 2010).

The KF approach has been demonstrated to be reliable (reliability of test scores for a 3.5 h exam containing 32 cases is about 0.70) and possess face and content validity in earlier studies (Farmer and Page 2005; Page and Bordage 1995). More recently, the KF approach was found to possess content evidence for validity as a self-assessment activity in surgical training (Trudel et al. 2008). Guidance on developing KF questions is available through the Medical Council of Canada. It is not entirely clear how KF type assessments fit into an overall assessment scheme and what they provide beyond single-best answer MCQs within an assessment program (Norman et al. 1996).

11.5.1.4 Script Concordance Testing

The script concordance test (SCT) was specifically developed to assess clinical reasoning that incorporates, to the extent possible, practice-like conditions (Dory et al. 2012). Unlike MCQs using extended matching or single-best answer, SCT is designed to produce a bounded range of responses based on clinical information provided sequentially to the trainee, in effect trying to simulate what might happen in a real clinical encounter. As such, the proponents of SCT argue that SCT may better address the uncertainty commonly encountered in real clinical practice (Charlin and van der Vleuten 2004). MCQs, extended matching and KF questions are not well designed to handle the ambiguity and complexity seen in day-to-day practice, especially when physicians face multiple competing pressures while caring for actual patients

In SCT, the learner is presented with a brief clinical scenario followed by questions that allow for the learner to list (or choose) several diagnostic or treatment options. Dory et al. (2012) provide an example in their recent review:

> A 25-year-old man presents to your general practice surgery. He has a severe retrosternal chest pain that began the previous night. There is nothing of note in his medical history. He does not smoke. His father, aged 60 years, and his mother, aged 55 years, are both in good health[1]

The possible response options are

If you were thinking of	And the patient reports or you find upon clinical examination	This hypothesis becomes				
Pericarditis	Normal chest auscultation	−2	−1	0	+1	+2
Pneumothorax	Decreased breath sounds in the left chest area with hyper-resonant chest percussion	−2	−1	0	+1	+2
Panic attack	Yellow deposits around the eyelids	−2	−1	0	+1	+2

Key *−2* ruled out or almost ruled out; *−1* less likely; *0* neither more nor less likely; *+1* more likely; *+2* certain or almost certain

After these initial answers, additional information is provided and the learner is asked how the new information will alter their diagnosis and/or management plan. Uncertainty can be built into the case to create a series of answers with a range of reasonable possibilities under the condition of uncertainty. SCT can also incorporate confidence ratings regarding the diagnosis. The learner’s answers are compared against the choices of an expert panel of at least 10–15 members that systematically generate a range of acceptable answers. For example, if 60 % of experts choose “−2” and 40 % choose “−1,” then examinees who choose “−2” receive full credit for the item (typically 1 point), and examinees who choose the latter option receive partial credit (typically 0.66 points, i.e., percentage of experts choosing minority option/percentage of experts choosing majority option, 40 %/60 %) (Dory et al. 2012). Bland et al. (2005) have highlighted that single-best answer scoring with three choices produces similar results to aggregate scoring, providing a potential option for simplifying grading.

[1] Adapted from Dory et al. (2012).

Lubarsky et al. (2013) recently published a useful AMEE Guide of SCT as well as a review of the current evidence for SCT. They concluded that some validity evidence exists for using SCT to assess clinical reasoning under the conditions of uncertainty and ambiguity, specifically the sequential interpretation of clinical information and data. However, it remains unclear what SCT adds to other assessment approaches for clinical reasoning and how SCT should be included in an assessment program and system (Lubarsky et al. 2013).

However, others have recently raised some legitimate concerns about the validity of the SCT approach. Lineberry et al. (2013, 2014) highlighted significant threats to the validity of SCT, including the assumption that all panelists' choices are equally valid and the lack of a valid psychometric model to account for its unique scoring aspects. Perhaps the most important contribution of the Lineberry et al. work is that traditional psychometric methods and theories may not be fully up to the task of more complex assessment methodologies and that much caution should be exercised in prematurely deploying new methods for routine use in assessing clinical reasoning.

11.6 Situated Assessment

11.6.1 Work-Based Assessments

11.6.1.1 "Expert" Assessments

"Expert" assessments are sometimes called "global summaries." They typically occur at the end of a learning experience (e.g., "a two-week rotation") and are "overall" or "summative" assessments performed by raters, typically faculty. These judgments have traditionally been expressed on some type of rating scale (e.g., a five-point Likert scale). While such ratings are imperfect and may provide minimal feedback for the learner, multiple studies have demonstrated positive correlations with faculty and program ratings and subsequent performance on certification exams (Haber and Avins 1994; Holmboe and Hawkins 1998; Norcini et al. 2013; Pangaro and Holmboe 2008).

Unfortunately, numerous limitations have been noted in the use of these scales for judging clinical reasoning, the most significant being poor inter-rater reliability ranging from 0.25 to 0.37 (Hawkins et al. 1999; Streiner 1985). Two factors that explain these poor correlations are faculty's clinical skills and lack of standardization of the clinical content and context of the reasoning case. Kogan et al. (2010) demonstrated that faculty clinical skills positively correlated with their rating stringency. Furthermore, raters often fail to appropriately recognize the important role and impact of contextual factors on the clinical reasoning process. Practically speaking, assessment programs attempt to overcome this limitation by obtaining multiple expert assessments over time for each learner.

While reviewing the health professions literature, we found one unique, validated expert-based assessment tool in nursing that merits mention, the Lasater Clinical Judgment Rubric (LCJR) (Lasater 2007, 2011). It is a promising developmental framework for evaluation of clinical judgment, which corresponds to clinical reasoning, with some validity and reliability data (Adamson et al. 2011), although it needs to be tested in larger and more diverse groups to confirm these preliminary findings. It expects an expert assessor to rate a learner's [what] on a developmental continuum (i.e., beginning to exemplary) in domains entitled noticing, interpreting, responding, and reflecting. The LCJR can be used for single or longitudinal observations.

11.6.1.2 Direct Observation

Direct observation has long been a mainstay method for the assessment of clinical skills. Although many tools exist, the Mini-Clinical Evaluation Exercise, or Mini-CEX is perhaps the most commonly used form (Kogan et al. 2009; Norcini et al. 2003). The original Mini-CEX specifically contains the rating domain of "clinical judgment." Assessing clinical judgment through observation commonly incorporates an assessment of data gathering skills (medical interviewing and physical examination), which is lacking in the other tools described. The advantage of direct observation, when combined with an assessment of data gathering skills, allows for faculty to assess integration of all the key components. For example, if the learner gathers data poorly, it substantially lowers the probability they can create an accurate problem representation and execute an appropriate treatment plan. Regarding the reasoning component, faculty can observe and assess for the construction of the problem representation, provide for supporting evidence (the "evaluation step") and assess the appropriateness of the action related to the decisions and reasoning around diagnosis (Gruppen et al. 1991).

While the Mini-CEX and other observation tools have been shown to be reliable and valid, much less is known about their psychometric properties specifically regarding assessment of clinical reasoning (Kogan et al. 2009). It would also be helpful to be able to tease out the various components of clinical reasoning on direct observation tools. Direct observation can also be combined with other assessments, such as assessing the ability to use evidence-based practice to answer clinical questions at the point of care (Holmboe 2004). Regardless of the direct observation tool, combining observation with effective use of questions or other adjuncts can provide meaningful feedback to learners.

11.6.1.3 Chart Stimulated Recall

Chart stimulated recall (CSR) can be considered as a more structured "oral exam" that uses the medical record of an actual patient encounter to retrospectively review the clinical reasoning process of healthcare professional (Maatsch et al. 1983). We recently referred to CSR as a form “game tape review” for health care providers (Holmboe and Durning 2014). Typically, the medical record of a clinical encounter, chosen by the health professional, trainee, or assessor, is first reviewed by the assessor against a structured template that produces a series of questions designed to probe the “why” behind the health professional’s actions and decisions (Chart Stimulated Recall, n.d.; PAR, n.d.; Schipper and Ross 2010). The assessor uses these questions in a one-on-one session with the health professional, eliciting and documenting the health professional’s rationale and reasoning for the choices reflected in the medical record, plus any additional pertinent information not documented. The challenge with such assessments, in addition to time requirements and training the assessors, is obtaining an adequate sampling of patient encounters and associated contexts. Sufficient sampling when using situated practice-based assessments is essential for high-stakes assessment.

CSR and a variant known as case-based discussion (CBD) are currently used in several contexts. CSR was originally developed for use in the American Board of Emergency Medicine certification in the early 1980s. While CSR was found to possess favorable psychometric properties, it was ultimately abandoned due to cost and the logistical difficulties (e.g., scheduling, recruiting sufficient faculty to perform the CSR) in operating a CSR-based high-stakes examination (Munger 1995; Munger et al. 1982). The main issues were the number of examiners needed and growing numbers of physicians entering emergency medicine. Currently, CSR is a validated component of the Physician Assessment Review (PAR) program in Canada (most notably the province of Alberta) and CBD has been studied as part of the United Kingdom’s Foundation program (Refs. 29–33 from Holmboe paper (Cunnington et al. 1997; Hall et al. 1999). Both techniques, used mostly for formative assessments, have been found to be reliable and perceived as useful by examiners and examinees alike. CSR has also been recently compared with chart audit among a group of family medicine physicians in Quebec. Agreement between CSR and chart audit in a limited sample for diagnostic accuracy was 81 %, but CSR predictably provided more useful information on clinical reasoning (Goulet et al. 2007).

11.6.2 Clinical Simulation-Based Assessments

In essence, tools such as key feature exams and script concordance testing (SCT) are forms of simulations around clinical reasoning through narrative and clinical data—however, they often do not involve interaction with any other physical material, such as equipment, clinical specimens, and other physical elements of the clinical setting where clinical encounters occur. Other forms of simulation can include the assessment of clinical reasoning that do combine either staged clinical interactions or incorporate other physical materials such as access to a computer and other medical equipment. We will discuss these in the following section.

11.6.2.1 OSCE and High Fidelity Simulations

The objective structured clinical examination (OSCE) is probably the most common format when standardized patients (SP) are employed to assess clinical skills. Standardized patients are live actors trained to portray a range of clinical scenarios. OSCEs are usually delivered as a series of stations where the learner is given 15–20 min to perform a focused medical history, physical examination, review of pertinent radiologic, or laboratory data with or without discussion or counseling with the SP. OSCES (Hawkins and Boulet 2008; Cleland et al. 2009). For example, both formative and summative formats often include an assessment of clinical reasoning (National Board of Medical Examiners, n.d.). United States Medical Licensing Examination (USMLE) step 2 clinical skills examination incorporates a patient note where the test taker provides a differential diagnosis. Similar techniques are also used in lower stakes OSCEs, including brief "oral exams" by faculty, interpretation of additional clinical material, and presentation of treatment plans to either the SP or faculty (Hawkins and Boulet 2008).

High fidelity simulation is also increasingly incorporating assessment of clinical reasoning into the assessment process. High fidelity simulations often do not involve an SP, but instead employ sophisticated mannequins, virtual reality, and other computer-based simulations. Like OSCEs, clinical reasoning through multiple approaches can be incorporated into the simulation (Scalese and Issenberg 2008; Walsh et al. 2012). Recently, the American Board of Anesthesiology (n.d.) added a simulation requirement to its maintenance of certification program, designed to help practicing anesthesiologists practice difficult and less common clinical scenarios in a controlled environment. Virtual reality is also enabling the creation of avatars that can be programmed to act like real patients and move down various pathways depending on the learner's clinical decisions (Satter et al. 2012; Courteille et al. 2008). What is less well-known is whether this type of assessment effectively transfers to actual clinical practice. Regardless, simulation holds significant promise as an assessment method.

11.7 Emerging Methods in Clinical Reasoning Assessment

A number of emerging means of assessing clinical reasoning are appearing in the literature. Here, we will briefly describe three such techniques recognizing that there are many others. The first two make an important advance with our assessment tools in that they explicitly incorporate educational theory. The latter shows promise by allowing for more direct introspection of the reasoning process.

Concept mapping is a technique for visually representing a learner's thinking or knowledge organization (Roberts 1999; Ruiz-Primo 2004; Schau and Mattern 1997). In a concept map, the learner connects a number of ideas (concepts) with specific phrases (linking words) to demonstrate how they put their ideas together. Concept maps can be unstructured (draw a concept map on the topic of anemia), partially structured (draw a concept map on the topic of anemia which includes hemoglobin, peripheral blood smear, microcytic, and bone marrow biopsy) to completely structured (whereby the learner fills in specific concepts or linking words, somewhat like a partially completed crossword puzzle). A number of recent publications on concept mapping have appeared in the medical education literature (Daley and Torre 2010; Torre et al. 2007).

Self-regulated learning (SRL) is defined as a set of processes that learners use to moderate their own learning and performance, which is typically divided into a number of elements in each of three stages: forethought (before), performance (during), and reflection (after) (Brydges and Butler 2012; Cleary et al. 2013; Durning et al. 2011; Zimmerman and Schunk 2011). These elements interact and from this model learning and performance is believed to emerge. More recently, medical education researchers have turned to theories of SRL—and, in particular, SRL microanalytic assessment techniques—to help understand and explain why and how some trainees succeed while others do not (Durning et al. 2011).

Neurobiological correlates are exploring clinical reasoning assessment through more direct means such as functional MRI and EEG (Durning et al. 2014). These means may show particular promise for nonanalytical reasoning that is not believed to be completely subject to introspection.

11.8 Conclusion

We have attempted to provide a working definition for clinical reasoning, describe the diagnostic reasoning process, and review multiple modalities for assessing clinical reasoning. While more assessment tools are needed, we can advance our theory and practice related to assessing clinical reasoning using the existing methods that we have and turning our focus to more authentic assessment of clinical reasoning. This will require neurobiological research into the clinical reasoning process and investment in faculty education to understand the clinical reasoning process, how to probe its key components, and how to leverage newer assessment methods to enhance clinical reasoning. In a competency-based world, assessment for learning is more important than assessment of learning. Such formative assessment will therefore need to focus more on clinically based tools like direct observation, meaningful simulation as starter and reinforcing activities, CSR, and new forms of audit.

There is a strong desire among educators to assist learners in developing good clinical reasoning, and the need for meaningful and effective remediation for those struggling with clinical reasoning. Practically speaking, how should health professions educators approach process assessment and provide formative feedback on clinical reasoning? Reflecting on the literature, we believe that the strongest recommendation that can be made is for educators to focus on helping learners build their discipline-specific knowledge and its organization (Elstein et al. 1978; Eva 2005). Given that much of the clinical reasoning process can be subconscious and is idiosyncratic (i.e., two health care professionals may come to the same conclusion using different processes based on their knowledge and experiences), educators must recognize that no "gold-standard" clinical reasoning process exists. In this relativistic world, knowledge assessment can provide a foundation. When a learner misses a diagnosis, the focus can first turn to the gaps in her knowledge (i.e., what knowledge was faulty or lacking that led her to the wrong diagnosis?).

As described in Sect. 11.3, deconstructing the clinical reasoning process into component parts can help focus the educator's assessment of a learner's knowledge. By exploring the components of reasoning, an educator may potentially help learners gain insight into how to improve their reasoning process (e.g., problem representation) if they encounter a similar case. For example, imagine a chest pain case where a student fails to mention a possible diagnosis of aortic dissection. The educator can ask the student to list the life threatening causes of chest pain and if the student misses dissection, then she can teach her the list of six deadly causes of chest pain and the typical presentation of aortic dissection. The educator should "probe" the student for knowledge rather than making assumptions about why a student missed a diagnosis. In addition, the literature supports the notion of using analytic approaches to confirm nonanalytic reasoning (Ark et al. 2007; Mamede et al. 2010). Students who seem to jump to conclusions about diagnosis or management can be encouraged to apply this process to their reasoning.

This chapter has sought to convey the complexities of clinical reasoning as a cognitive process, as a clinical performance, and as a target for assessment. We want to emphasize that no one of the several assessment methods described above is adequate as a measure of clinical reasoning. Instead, it may be helpful to consider the old Indian story of the blind men and the elephant, in which each blind man, grasping a different part of the elephant, described this beast as similar to a palm leaf (the ear), a tree (the leg), a snake (the tail), a wall (the body), or a tree branch (the trunk). Only by bringing together all of the diverse descriptions can we obtain a complete understanding of this thing called "clinical reasoning."

Issues/Questions for Reflection

- How do we define clinical reasoning competence given the issue of context specificity?
- How do we assess clinical reasoning under conditions of diagnostic uncertainty given that there is no gold standard for assessing the clinical reasoning process?
- Are there ways to improve inter-rater reliability as it relates to clinical reasoning assessment?
- Can neurobiological assessments (e.g., functional MRI or EEG) provide greater insights into the diagnostic process and improve our ability to assess clinical reasoning?

References

Abbott, V., Black, J. B., & Smith, E. E. (1985). The representation of scripts in memory. *Journal of Memory and Language, 24*(2), 179–199.

ACP Smart Medicine. (n.d.). Retrieved July 28, 2014, from http://smartmedicine.acponline.org

Adamson, K. A., Gubrud, P., Sideras, S., & Lasater, K. (2011). Assessing the reliability, validity, and use of the Lasater Clinical Judgment Rubric: three approaches. *Journal of Nursing Education, 51*(2), 66–73.

American Board of Anesthesiology. (n.d.). *Maintenance of certification in anesthesiology (MOCA): Simulation for MOCA.* Retrieved July 22, 2014 from http://www.theaba.org/Home/anesthesiology_maintenance

Ark, T. K., Brooks, L. R., & Eva, K. W. (2007). The benefits of flexibility: The pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. *Medical Education, 41*(3), 281–287.

Babbott, S. F., Beasley, B. W., Hinchey, K. T., Blotzer, J. W., & Holmboe, E. S. (2007). The predictive validity of the internal medicine in-training examination. *American Journal of Medicine, 120*(8), 735–740.

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine*, *80*(4), 395–399.

Bordage, G. (2007). Prototypes and semantic qualifiers: From past to present. *Medical Education, 41*(12), 1117–1121.

Bordage, G., & Page, G. (2012, August). Guidelines for the development of key feature problems and test cases. *Medical Council of Canada.* Retrieved July 20, 2014, from http://mcc.ca/wp-content/uploads/CDM-Guidelines.pdf

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Research, 18*(1), 32–42.

Brydges, R., & Butler, D. (2012). A reflective analysis of medical education research on self-regulation in learning and practice. *Medical Education*, *46*(1), 71–79.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine: An International Journal, 5*(2), 107–115.

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd Ed.). National Board of Medical Examiners. Retrieved February 6, 2015, from http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf

Chang, D., Kenel-Pierre, S., Basa, J., Schwartzman, A., Dresner, L., Alfonso, A. E., & Sugiyama, G. (2014). Study habits centered on completing review questions result in quantitatively higher American Board of Surgery In-Training Exam scores. *Journal of Surgical Education, 71*(6), e127–e131.

Charlin, B., Boshuizen, H., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education, 41*(12), 1178–1184.

Charlin, B., & van der Vleuten, C. (2004). Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. *Evaluation and the Health Professions, 27*(3), 304–319.

Chart Stimulated Recall. (n.d.). *Practical Doc: By rural doctors, for rural doctors.* Retrieved February 2, 2015, from http://www.practicaldoc.ca/teaching/practical-prof/teaching-nuts-bolts/chart-stimulated-recall/

Cleary, T. J., Durning, S. J., Gruppen, L. D., Hemmer, P. A., & Artino Jr, A. R. (2013). Self-regulated learning. *Oxford textbook of medical education*, 465–478.

Cleland, J. A., Abe, K., & Rethans, J. J. (2009). The use of simulated patients in medical education: AMEE Guide No 42 1. *Medical Teacher, 31*(6), 477–486.

Counselman, F. L., Borenstein, M. A., Chisholm, C. D., Epter, M. L., Khandelwal, S., Kraus, C. K., et al. (2014). The 2013 model of the clinical practice of emergency medicine. *Academic Emergency Medicine, 21*(5), 574–598.

Courteille, O., Bergin, R., Courteille, O., Bergin, R., Stockeld, D., Ponzer, S., & Fors, U. (2008). The use of a virtual patient case in an OSCE-based exam-a pilot study. *Medical Teacher, 30*(3), e66–e76.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775–780.

Cunnington, J. P., Hanna, E., Turnhbull, J., Kaigas, T. B., & Norman, G. R. (1997). Defensible assessment of the competency of the practicing physician. *Academic Medicine, 72*(1), 9–12.

Daley, B. J., & Torre, D. M. (2010). Concept maps in medical education: an analytical literature review. *Medical Education, 44*(5), 440–448.

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: Insights from a systematic review. *Medical Education, 46*(6), 552–563.

Durning, S. J., & Artino, A. R. (2011). Situativity theory: A perspective on how participants and the environment can interact: AMEE Guide no. 52. *Medical Teacher, 33*(3), 188–199.

Durning, S. J., Artino, A. R, Jr, Schuwirth, L., & van der Vleuten, C. (2013). Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. *Academic Medicine, 88*(4), 442–448.

Durning, S. J., Cleary, T. J., Sandars, J., Hemmer, P., Kokotailo, P., & Artino, A. R. (2011). Perspective: Viewing “strugglers” through a different lens: How a self-regulated learning perspective can help medical educators with assessment and remediation. *Academic Medicine, 86*(4), 488–495.

Durning, S. J., Costanzo, M., Artino, A. R., Vleuten, C., Beckman, T. J., Holmboe, E., et al. (2014). Using functional magnetic resonance imaging to improve how we understand, teach, and assess clinical reasoning. *Journal of Continuing Education in the Health Professions, 34* (1), 76–82.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Medical Education, 41*(12), 1124–1130.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, UK: Cambridge University Press.

Eva, K. W. (2003). On the generality of specificity. *Medical Education, 37*(7), 587–588.

Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical Education, 39*(1), 98–106.

Eva, K. W., Hatala, R. M., LeBlanc, V. R., & Brooks, L. R. (2007). Teaching from the clinical reasoning literature: Combined reasoning strategies help novice diagnosticians overcome misleading information. *Medical Education, 41*(12), 1152–1158.

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Academic Medicine, 73*(10), S1–S5.

Farmer, E. A., & Page, G. (2005). A practical guide to assessing clinical decision-making skills using the key features approach. *Medical Education, 39*(12), 1188–1194.

Fonteyn, M., & Grobe, S. (1993). Expert critical care nurses’ clinical reasoning under uncertainty: Representation, structure and process. In M. Frisee (Ed.), *Sixteenth annual symposium on computer applications in medical care* (pp. 405–409). New York, NY: McGraw-Hill.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*(10), S1–S7.

Goulet, F., Gagnon, R., & Gingras, M. É. (2007). Influence of remedial professional development programs for poorly performing physicians. *Journal of Continuing Education in the Health Professions, 27*(1), 42–48.

Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education, 16*(2), 151–165.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493–1499.

Grabovsky, I., Hess, B. J., Haist, S. A., Lipner, R. S., Hawley, J. L., Woodward, S., et al. (2014). The relationship between performance on the infectious disease in-training and certification examinations. *Clinical Infectious Diseases*, ciu906v2.

Green, M. L., Reddy, S. G., & Holmboe, E. (2009). Teaching and evaluating point of care learning with an Internet-based clinical-question portfolio. *Journal of Continuing Education in the Health Professions, 29*(4), 209–219.

Gruppen, L. D., & Frohna, A. Z. (2002). Clinical reasoning. In G. R. Norman, C. P. M. van der Vleuten, & D. I. Newble (Eds.), *International handbook of research in medical education* (pp. 205–230). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Gruppen, L. D., Wolf, F. M., & Billi, J. E. (1991). Information gathering and integration as sources of error in diagnostic decision making. *Medical Decision Making, 11*(4), 233–239.

Haber, R. J., & Avins, A. L. (1994). Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *Journal of General Internal Medicine, 9*(3), 140–145.

Hall, W., Violato, C., Lewkonia, R., Lockyer, J., Fidler, H., Toews, J., & Moores, D. (1999). Assessment of physician performance in Alberta the physician achievement review. *Canadian Medical Association Journal, 161*(1), 52–57.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81–112.

Hawkins, R. E., & Boulet J. R. (2008). Direct observation: Standardized patients. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp.102–118). Philadelphia, Pa: Elsevier.

Hawkins, R. E., Lipner, R. S., Ham, H. P., Wagner, R., & Holmboe, E. S. (2013). American board of medical specialties maintenance of certification: Theory and evidence regarding the current framework. *Journal of Continuing Education in the Health Professions, 33*(S1), S7–S19.

Hawkins, R. E., Sumption, K. F., Gaglione, M. M., & Holmboe, E. S. (1999). The in-training examination in internal medicine: Resident perceptions and lack of correlation between resident scores and faculty predictions of resident performance. *The American Journal of Medicine, 106*(2), 206–210.

Hodges, B. D. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568.

Holmboe, R. S. (2004). Tbc importance of faculty observation of trainees' clinical skills. *Academic Medicine, 79*, 16–22.

Holmboe, E. S., & Durning, S. J. (2014). Assessing clinical reasoning: Moving from in vitro to in vivo. *Diagnosis, 1*(1), 111–117.

Holmboe, E. S., & Hawkins, R. E. (1998). Methods for evaluating the clinical competence of residents in internal medicine: A review. *Annals of Internal Medicine, 129*(1), 42–48.

Holmboe, E. S., Lipner, R., & Greiner, A. (2008). Assessing quality of care: Knowledge matters. *JAMA, 299*(3), 338–340.

Jefferies, A., Simmons, B., Ng, E., & Skidmore, M. (2011). Assessment of multiple physician competencies in postgraduate training: Utility of the structured oral examination. *Advances in Health Sciences Education Theory and Practice, 16*(5), 569–577.

Johnson, E. J., Camerer, C., Sen, S., & Rymon, T. (1991). *Behavior and cognition in sequential bargaining*. Wharton School, University of Pennsylvania, Marketing Department.

Jones, M. A., Jensen, G., & Edwards, I. (2008). Clinical reasoning in physiotherapy. In. J. Higgs, M. A. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical reasoning in the health professions* (3rd Ed., pp. 245–256). New York: Elsevier Limited.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus, & Giroux.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772–775.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Kogan, J. R., Hess, B. J., Conforti, L. N., & Holmboe, E. S. (2010). What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Academic Medicine, 85* (10), S25–S28.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*(12), 1316–1326.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory and Practice, 414*(4), 212–218.

Larsen, D. P., Butler, A. C., & Roediger, H. L, I. I. I. (2008). Test-enhanced learning in medical education. *Medical Education, 42*(10), 959–966.

Lasater, K. (2007). Clinical judgment development: Using simulation to create an assessment rubric. *Journal of Nursing Education, 46*(11), 496–503.

Lasater, K. (2011). Clinical judgment: The last frontier for evaluation. *Nurse Education in Practice, 11*(2), 86–92.

Lave, J. (1988). *Cognition in practice*. Cambridge, MA: Cambridge University Press.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education, 47*(12), 1175–1183.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2014). Script concordance tests: Strong inferences about examinees require stronger evidence. *Medical Education, 48*(4), 452–453.

Liu, K. P., Chan, C. C., & Hui-Chan, C. W. (2000). Clinical reasoning and the occupational therapy curriculum. *Occupational Therapy International, 7*(3), 173–183.

Lubarsky, S., Dory, V., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script concordance testing: From theory to practice: AMEE Guide No. 75. *Medical Teacher, 35*(3), 184–193.

Maatsch, J. L., Huang, R., Downing, S. M., & Barker, D. (1983). Predictive validity of medical specialty examinations. Final report to NCHSR Grant No.: HS02039-04.

Mamede, S., Schmidt, H. G., Rikers, R. M., Custers, E. J., Splinter, T. A., & van Saase, J. L. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research, 74*(6), 586–592.

McCarthy, W. H., & Gonnella, J. S. (1967). The simulated patient management problem: A technique for evaluating and teaching clinical competence. *Medical Education*, *1*(5), 348–352.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67.

Munger, B. S. (1995). Oral examinations. *Recertification: New Evaluation Methods and Strategies* (pp. 39–42). Evanston, ILL: American Board of Medical Specialties.

Munger, B. S., Krome, R. L., Maatsch, J. C., & Podgorny, G. (1982). The certification examination in emergency medicine: An update. *Annals of Emergency Medicine, 11*(2), 91–96.

National Board of Medical Examiners International Foundations of Medicine. (n.d.). Retrieved July 25th, 2014, from http://www.nbme.org/ifom/

National Board of Medical Examiners. (n.d.). *Step 2 clinical skills*. Retrieved July 22, 2014, from http://www.usmle.org/pdfs/step-2-cs/cs-info-manual.pdf

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., & Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(3), 206–214.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476–481.

Norcini, J. J., Lipner, R. S., & Grosso, L. J. (2013). Assessment in the context of licensure and certification. *Teaching and Learning in Medicine, 25*(Suppl1), S62–S67.

Norcini, J. J., Swanson, D. B., Grosso, L. J., Shea, J. A., & Webster, G. D. (1984). A comparison of knowledge, synthesis, and clinical judgment multiple-choice questions in the assessment of physician competence. *Evaluation and the Health Professions, 7*(4), 485–499.

Norcini, J. J., Swanson, D. B., & Webster, G. D. (1982). Reliability, validity and efficiency of various item formats in assessment of physician competence. In Proceedings of the Annual

Conference on Research in Medical Education. Conference on Research in Medical Education (Vol. 22, pp. 53–58).

Norman, G. R., Swanson, D. B., & Case, S. M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine: An International Journal, 8*(4), 208–216.

Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine, 70*(2), 104–110.

Pangaro, L., & Holmboe, E.S. (2008). Evaluation forms and rating scales. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 102–118). Philadelphia, PA: Mosby-Elsevier.

Pauker, S. G., Gorry, G. A., Kassirer, J. P., & Schwartz, W. B. (1976). Towards the simulation of clinical cognition: Taking a present illness by computer. *The American Journal of Medicine, 60* (7), 981–996.

Physician Achievement Review (PAR). (n.d.). Retrieved July 20, 2014, from http://parprogram.org/par/

Rimoldi, H. J. (1963). Rationale and applications of the test of diagnostic skills. *Academic Medicine, 38*(5), 364–368.

Roberts, L. (1999). Using concept maps to measure statistical understanding. *International Journal of Mathematical Education in Science and Technology, 30*(5), 707–717.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Potomac, MD: Erlbaum Press.

Ruiz-Primo, M. A. (2004). Examining concept maps as an assessment tool. In A. J. Canas, J. D. Novak, & F. M. Gonzalez (Eds.), *Concept maps: Theory, methodology, technology. Proceedings of the First International Conference on Concept Mapping* (pp. 555–562). Pamplona, Spain.

Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge, UK: Cambridge University Press.

Satter, R. M., Cohen, T., Ortiz, P., Kahol, K., Mackenzie, J., Olson, C., & Patel, V. L. (2012). Avatar-based simulation in the evaluation of diagnosis and management of mental health disorders in primary care. *Journal of Biomedical Informatics, 45*(6), 1137–1150.

Scalese, R. S., Issenberg, S. B. (2008). Simulation-based assessment. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence.* Philadelphia: Mosby-Elsevier.

Schau, C., & Mattern, N. (1997). Use of mapping techniques in teaching applied statistics courses. *The American Statistician, 51*, 171–175.

Schipper, S., & Ross, S. (2010). Structured teaching and assessment. *Canadian Family Physician, 56*(9), 958–959.

Schmidt, H. G., & Rikers, R. M. (2007). How expertise develops in medicine: Knowledge encapsulation and illness script formation. *Medical Education, 41*(12), 1133–1139.

Schuwirth, L. (2009). Is assessment of clinical reasoning still the Holy Grail? *Medical Education, 43*(4), 298–300.

Schuwirth, L. W. T., & van der Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*, 296–300.

Streiner, D. L. (1985). Global rating scales. In: V. R. Neufeld & G. R. Norman (Eds.), *Assessing clinical competence* (pp. 119–141). New York, NY: Springer.

Sweller, J., Van Merrienböer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Torre, D. M., Daley, B., Stark-Schweitzer, T., Siddartha, S., Petkova, J., & Ziebert, M. (2007). A qualitative evaluation of medical student learning with concept maps. *Medical Teacher, 29* (9–10), 949–955.

Trudel, J. L., Bordage, G., & Downing, S. M. (2008). Reliability and validity of key feature cases for the self-assessment of colon and rectal surgeons. *Annals of Surgery, 248*(2), 252–258.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

van Der Vleuten, C. P. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41–67.

van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317.

van Merrienböer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review, 17*(2), 147–177.

Walsh, C. M., Sherlock, M. E., Ling, S. C., & Carnahan, H. (2012). Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy. *Cochrane Database of Systematic Reviews, 6*, 1–91. doi:10.1002/14651858.CD008237.pub2

Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. In: B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 49–64). New York, NY: Routledge.

Chapter 12
Assessment of Interprofessional Education: Key Issues, Ideas, Challenges, and Opportunities

Brian S. Simmons, Susan J. Wagner and Scott Reeves

Abstract Over the past few decades interprofessional education (IPE) has grown within the health professional education. IPE aims to provide learners with interactive experiences in order to prepare them better to work collaboratively to effectively meet the needs of patients, clients, and families. While the IPE literature has expanded significantly in the past few years, efforts to produce rigorous assessment of IPE continue to be a challenge. At present, most IPE assessment is focused on learner self-assessment that only provides a perception of what the learner thinks she/he may have learned. These struggles with assessing IPE appear to be rooted in a number of factors, including uncertainty about what to assess (e.g., individuals, groups, and/or teams), logistical challenges with organizing assessment for large groups of students and limited resources for IPE assessment. Despite these difficulties, it is recommended that the principles of assessment should be adhered to in any IPE activity. This chapter provides an exploration of key issues related to the assessment of IPE. It outlines the purpose of assessment and the use of an assessment blueprint to ensure that learners cover the relevant collaborative competencies. It also reflects on the processes of designing and implementing an IPE assessment activity and conceptualizes what needs to be assessed in IPE. This is illustrated using a clinical competency continuum model that employs the concept of milestones and applies 'entrusted professional activities' in a performance framework.

B.S. Simmons (✉)
Department of Paediatrics, University of Toronto, Toronto, ON, Canada
e-mail: brian.simmons@sunnybrook.ca; brain.simmons@uhn.org

S.J. Wagner
Department of Speech-Language Pathology, Faculty of Medicine,
University of Toronto, Toronto, ON, Canada

S. Reeves
Faculty of Health, Social Care and Education, Kingston University,
London, UK
e-mail: s.reeves@sgul.kingston.ac.uk

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_12

Takeaways

- The assessment of interprofessional education (IPE) is a complex activity due to its involvement of both individuals as well as interprofessional teams/groups.
- In designing an IPE assessment a series of key questions need to be posed and addressed, including, what is the purpose of the assessment? What is one going to assess? How is the assessment to be performed?
- Development of assessment blueprint is vital to linking proposed learning outcomes with methods of assessment.
- A focus on collaborative performance using competency domains such as communication, collaboration and professionalism can be an effective approach to IPE assessment.
- The use of an assessment matrix can effectively collate key elements related to the assessment of IPE.
- Entrustable professional activities and milestones are promising techniques to use in IPE assessment.

12.1 Introduction

IPE focuses on learning activities designed to enhance the attitudes, knowledge, skills, and behaviors for effective interprofessional practice (Barr et al. 2005). Recently, these attributes have been developed and categorized into 'IPE competencies' by the national bodies advocating the use of this form of education (Interprofessional Education Collaborative Expert Panel 2011; Canadian Interprofessional Health Collaborative 2010). In the US, the Interprofessional Education Collaborative Expert Panel (2011) has created of IPE competencies which have centered on the use of key domains: Values/Ethics for Interprofessional Practice; Roles/Responsibilities; Interprofessional Communication; and Teams and Teamwork. Within each of these domains, there are a series of general competency statements and also specific competencies statements that support the development of competencies linked to enhancing collaboration and teamwork under each domain.

Through the use of IPE, it is anticipated that improvements in the quality of care delivered to patients/clients and families will be achieved (e.g., Reeves et al. 2010, 2013; Institute of Medicine 2013, 2014). These aims were reemphasized in two significant policy reports. The first, from the World Health Organization (2010), outlined the role and attributed impact of IPE in preparing health care providers to enter the workplace as a member of the collaborative practice team. These sentiments were echoed in a second policy document that called for the use of IPE to promote effective collaborative care (Frenk et al. 2010).

As a result, IPE activities are increasingly being offered to a range of learners from different groups of health science professions (e.g., Pollard and Miers 2008; Lundon et al. 2013). This growth in IPE activities has, in turn, led to an expansion of the literature (e.g., Lidskog et al. 2009; Curran et al. 2011; Reeves et al. 2013). Systematic reviews of the literature, primarily of IPE evaluation studies, suggest that IPE can have a beneficial impact on learners' ability to work together in an effective manner collaborative attitudes, knowledge, skills, and behaviors (Zwarenstein et al. 1999; Hammick et al. 2007; Reeves et al. 2013).

While the evaluation of IPE programs continues to grow, in contrast the assessment[1] of learning in IPE has received far less consideration, with only a limited amount of literature published (e.g., Morison and Stewart 2005; Institute of Medicine 2014). Moreover, there currently is very limited evidence about what learning occurs during an IPE activity. Typically, evidence of the impact of IPE is generated from evaluating IPE programs and its utility, often linked to self-reported changes in knowledge and skills.

The current shortfall in relation to assessing IPE centers on how to design and implement rigorous assessment approaches for this type of education. Key questions to consider include: should one use a summative approach to assessing learning in interprofessional groups or teams or is a formative assessment approach more effective or should both be utilized? What should be the focus of the assessment: the individual, the team or the completion of the task? What should one measure: individual-based, patient/client centered-based, organizational-based outcomes or others?

In this chapter an exploration of key issues related to IPE assessment is provided. Core concepts and principles of good assessment are presented. This is followed by our experiences in designing and introducing an IPE assessment activity into an existing university IPE curriculum using a blueprint to ensure coverage of relevant collaborative competencies. As well, a conceptualization is offered of what needs to be assessed in IPE from the individual performance to the function of the team, and to achieving completion of agreed collaborative tasks.

12.2 Assessment Development: Key Principles

The development of an IPE assessment involves posing and then responding to the following questions: what is the purpose of the assessment? What is one going to assess? Who is going to be assessed? And how is the assessment to be performed? Also, in designing an assessment one needs to ask: what does assessment offer the learner and what recognition (i.e., certificates or diplomas) will be given for their involvement in IPE?

[1] Assessment is defined as a measurement of learning for 'people' or individuals alone or together in group or teams. It contrasts with the term evaluation that is the measurement of 'things' such as educational programs, interventions or courses (Institute of Medicine 2013).

Competency frameworks which can map learners' transition through different stages—from novice, advanced beginner, competence, proficiency, and finally expert (Dreyfus 1972; Benner 1982) are useful to employ. The assessment of such a competency-based approach in IPE may help to determine if and when such desired progress has occurred. For example, whether, and to what extent, IPE programs are producing clinicians who are able to collaborate effectively in interprofessional teams and groups.

12.2.1 Collaborative Performance

Within an IPE program or curriculum, working together collaboratively is the core issue. However, determining which elements constitute an 'adequate' collaborative performance is a complex activity. The difficulties with determining adequate performance or competence can be illustrated by drawing upon some of our work in creating an IPE curriculum for students in eleven health science programs at the University of Toronto, Canada. At present, there are differences in the examination standards and assessment criteria required across the eleven profession-specific programs. In introducing an IPE curriculum, agreement was needed for both examination standards and assessment criteria—these should be equivalent across all programs. As these assessment issues are complex, it was not surprising that the leaders of the different programs initially agreed on a formative approach to assessment. While this discussion on assessment ensured that an IPE curriculum could begin to be developed, however, the assessment of collaborative performance was limited, and often not regarded as meaningful as the use of summative assessment. In addition, each professional program determined what is called 'successful completion', which did result in difficulties in consistently determining this outcome. However, at present, unsuccessful completion of the IPE program does not affect a student's ability to graduate from their program.

These assessment problems can also be seen at the national level. For example, professional regulatory bodies such as the Royal College of Physicians and Surgeons of Canada have determined the collaborative competencies for their own professional group (Royal College of Physicians and Surgeons of Canada 2005). While, however, there are some similarities among profession-specific competency-based approaches, there is generally little agreement to produce one shared competency-based framework, resulting in an on-going uncertainty about how to assess collaborative team performance. Each profession is, therefore, determining what it constitutes to be a collaborative competency-based team. Although the development of IPE competency frameworks by groups such as the Canadian Interprofessional Health Collaborative (2010) and the Interprofessional Education Collaborative Expert Panel (2011) have produced a set of common competencies, there is a good deal of uncertainty in how one assesses interprofessional performance (Reeves 2012).

12.2.2 What to Assess

Regardless of the standards and criteria, it is useful for IPE programs to know upon what learning activities they are being assessed. Within each learning activity, assessment must include knowledge, application of knowledge, performance of the knowledge, and what we do in reality which would be to develop a professional competence in practice: ‘knows’; ‘knows how’; ‘shows how’, and ‘does’ (Miller 1990). This approach can also be defined as content-specific assessment or domain-specific assessment (van der Vleuten 2008). As learners move from ‘knows’ (knowledge) to ‘does’ (performance), there is a need to assess progression from cognitive to behavioral abilities. However, each level is dependent on the other and also dependent upon content. The competencies for IPE are domain independent outcomes and are related to each level of Miller’s typology and independent of content. It is important that the team develops core competencies in these domains this should be related to structure of the team (the individual roles) function (team roles) and outcome (task completion).

12.2.3 Development of a Blueprint

A blueprint links learning outcomes with methods of assessment, the target phase of learning in which the learning outcome should be achieved and also maps the assessment to practice. Using a blueprint for IPE can ensure a range of competencies (e.g., ethics/values, roles/responsibilities) are used to cover all proposed learning activities. An IPE blueprint can also help in identifying how to assess the different competencies, which can provide more balance in the blueprint (Hamdy 2006; D’Eon 2004). A balanced assessment blueprint would include a range of methodologies, such as self-assessment; peer assessment; case-based assessment, and/or team assessment.

12.2.4 Assessing Collaborative Performance

Although, as discussed above, assessing individual learners in individual learning activities is a challenge in IPE, the ultimate aim must be to assess collaboration and teamwork. Attempting to assess collaboration can be problematic, especially when a group of learners are brought together, often without preparation, and required to perform together as an interprofessional team. In practice, however, interprofessional teams may have worked together for many years. Nevertheless, team structure, size, and composition are varied with different levels of professional and interprofessional experience and expertise.

Methodologies to assess interprofessional teamwork are limited in nature, with most relying upon the use of self-assessment (Reeves et al. 2010). One helpful approach to understanding the nature of team development is Tuckman's (1965) work. He maintained that teams typically transition through the following stages: 'forming', 'storming', 'norming', 'performing', and finally 'adjourning'. The time taken to move through and among these stages of development will depend upon a number of individual and organizational factors (e.g., understanding of roles and responsibilities, management support for interprofessional collaboration). It also assumes that all team members enter the team task environment at the same level of competency—which is unlikely. One potentially useful approach to assess interprofessional team performance in a standard manner, given the complexities outlined above, is to employ an objective structured clinical examination (OSCE).

Symonds et al. (2003) describe the use of an OSCE—in the form of an interprofessional team objective structured clinical examination (ITOSCE)—by employing a mixed group of medical and midwifery learners rotating through a series of scenarios on common labor room problems. Five team examination stations were developed using a checklist related to the task and teamwork. Feedback related to problem-solving skills, knowledge, and attitude to team working was reported to the learners by a facilitator. Symonds and colleagues' note, however, that their ITOSCE was a formative exercise where the occurrence of actual learning was not determined that was also resource intensive and logistically challenging to implement. It had limited domains for assessment and; therefore, formative feedback to the team was very appropriate as assessment of individuals or team within the task was not possible as there was limited validity/reliability as per the design of the study. Also, using a traditional checklist a formative OSCE does not determine expertise along the clinical competency continuum (Chumley 2008).

In order to undertake a team-OSCE (TOSCE), learners must be given time to 'form' in order to perform the task. The time period required to 'form' varies from team to team and; therefore, has not yet been determined. Arguably, 'forming' is dependent upon an understanding of roles, responsibilities and relationships. Part of the 'forming' is a briefing or planning process to determine the roles that occur within the task requested. This will require an ability to self-reflect or use self-awareness skills that will need to be developed prior to the team assessment process (Singleton et al. 1999). Although in well-established teams forming is not an issue, the introduction of a new member(s) means that effort is again needed at this stage to integrate the new member(s).

Schön (1987) described how 'reflection in action' and 'reflection on action' are crucial steps in the learning process of individuals. This of course must be true of teams and if we add in the additional step of 'reflection before action' this matches the process of team forming/norming, task performing, and team adjourning as above.

To develop a five-station or scenario TOSCE simulation with four to five learners per station and a scenario length of 40–45 min will require a 4-h process with a total of 20–25 learners. Given these needs, the content of the scenarios would be limited and; therefore, would diminish the validity, reliability (dependent upon adequate rater training) and acceptability due to cost. Although the practicability of the TOSCE may be limited, it has a significant amount of potential as a teaching tool based on feedback and reflection. It also has potential in terms of educational impact, although this is likely to be formative rather than summative.

As can be seen, the assessment of learners in IPE is fraught with conceptual and practical difficulties. Program planners and designers must be aware of outcomes that are the collaborative competencies, as well as design blueprints for both the programs and assessments. They also need to use an assortment of methods and tools to ensure that learning has occurred in an assortment of domains. This has to be related to both content and process to increase the authenticity of IPE processes. This must be done for both individuals and the interprofessional teams within which those individuals collaborate together.

12.2.5 Designing and Piloting an Interprofessional Assessment

Based on this literature, an interprofessional OSCE (iOCSE) was developed at the University of Toronto, Canada in which we designed a three-station assessment with five students from different professions (Simmons et al. 2012). Each scenario consisted of a briefing or planning session, the task and a debriefing session. Assessors or raters were from three different health science professions. Students were given background information for the briefing/planning session where they were able to begin to form, storm, and norm (Tuckman 1965). The task was a live interaction with one or two standardized patients, clients, and/or families where the team was able to perform. Debriefing consisted of a session with the standardized patients and with the assessors.

Content for this assessment activity was initially determined through a modified Delphi consensus-building approach with IPE experts from different professions (Simmons et al. 2011). Although difficult to assess the professions individually, there was progression across the clinical competency continuum from novice to competence through the three scenarios. Evaluations of these activities indicated that students valued this interprofessional experience. As students had not had the opportunity to interact interprofessionally before, the briefing session was inadequate time for the team to determine the individual roles and responsibilities of each profession and was one reason team expertise evolved through the scenarios due to the fact students were able to get to know other professions in more detail.

12.2.6 A New Approach to Assessing Interprofessional Education

Based on these experiences with assessing IPE, in this section a new possible approach to IPE assessment is described in which a range of interlinked concepts and activities is explored.

12.2.6.1 Structure-Function-Outcome

Perhaps assessment in IPE requires looking at the task in a different way and using a different approach. Clinicians take many images of patients (e.g., X-rays, MRI scans) that are representations of ‘structure’ only. As such these images provide no insights into how the patient feels, the nature of the clinicians’ work (function) or the effect of their work on clinical outcomes.

We can apply this notion of ‘structure-function-outcomes’ to IPE. In observing an interprofessional team learning together to achieve success in a clinical task, their interactions typically illustrate only an image of collaboration (among individuals), from which function (team interactions) or potential outcomes (task completion) are not possible to determine.

Figure 12.1 provides a representation of a structure-function-outcomes approach to the assessment of IPE that focuses on individuals, teams, and tasks. Given the different types of activities involved in the assessment of IPE at the individual, team, and task levels, the methods used for assessment would be different in all three.

12.2.6.2 Individuals-Team-Task

Given that these links are not in reality linear with interprofessional teams, Fig. 12.2 more accurately conveys a sense of the interplay among structure, function, outcome, and correspondingly, individual, team, and task.

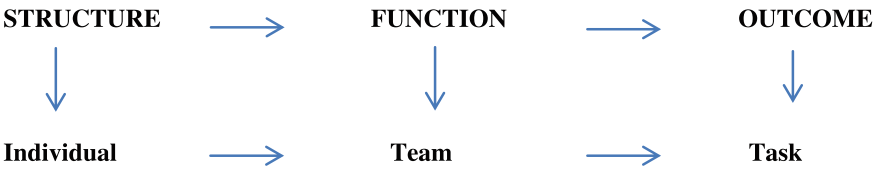


Fig. 12.1 Structure-function-outcomes model for assessment (linear)

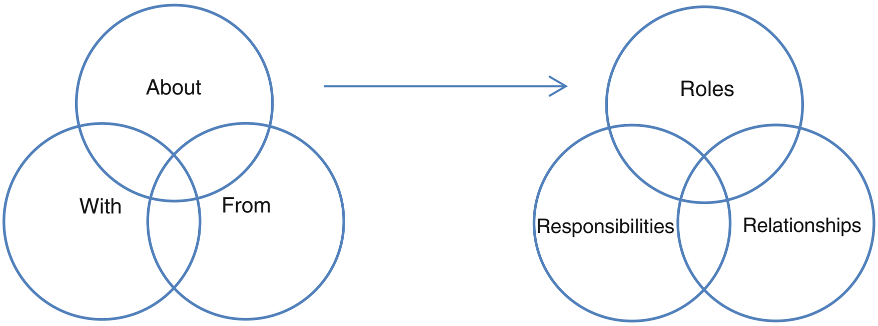


Fig. 12.2 Structure-function-outcomes model for assessment (integrated)

12.2.6.3 Focusing the Assessment

To assess all three domains in IPE (individual, team, and task) together and simultaneously in relation to structure, function and outcome is an extremely difficult task. However, if assessment of IPE is divided according to the well-known definition of IPE, "learning about, from and with different professions to improve collaboration and patient care" (Centre for the Advancement of Interprofessional Education—CAIPE 2002), this then takes the assessment activity a step further. Using such an approach, an IPE activity (learning about, from and with) could be linked to the learner understanding roles, responsibilities, and relationships involved in interprofessional collaboration—see Fig. 12.3.

12.2.6.4 The Use of a Matrix

The concepts discussed above can be usefully collated into an assessment matrix under the four headings of 'who', 'how', 'what', and 'assessment' (see Table 12.1). Also included in the table are examples of possible assessment tools. As noted above, assessment of IPE is multidimensional. Individuals, as members of interprofessional teams, through their interrelationships with other team members effectively form the team structure. However, it is unlikely that a team will be able to function effectively if the individual professions are unaware of what contributions their colleagues from other professions make. In order for the team to

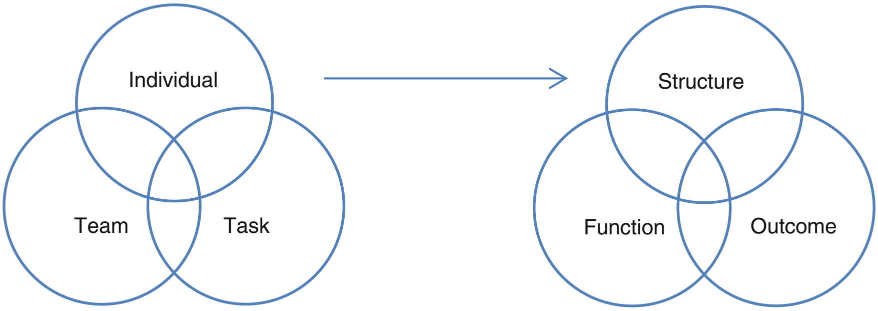


Fig. 12.3 Integrated approach to the definition of IPE as applied to assessment

Table 12.1 Assessment matrix

Who	How	What	Assessment	Example
Individual	Structure	Profession	Knows	MCQ, SAQ, self-assessment
Team	Function	Competence	Knows how	PBL, On-line modules
Task	Outcome	Performance	Shows how	ITOSCE, OSCE

function, the professional roles of the different individuals should be understood by all team members. This can then allow team members to be aware of the competencies expected of them. As a result, the task can be completed based upon a collaborative performance that can incorporate structure linked to individual professions and function of the team/group. Therefore, to function as an effective interprofessional team we ought to assess structure before the task is completed. The structure and function must be assessed to determine the level of attainment to enable completion of the task.

12.2.6.5 Mapping Assessment Options: An Interprofessional Blueprint

Most academic institutions have multiple health and social care professionals who will be undertaking IPE together. Should all these professionals be aware of the roles responsibilities and relationships of the other professions? The reality is that not all professionals will be on a team at the same time. This takes us back to structure. When interprofessional teams of learners are aware of their composition (structure), IPE learning activities need to be created that aim to develop their understanding of their different professional roles, responsibilities and relationships within the team. The development of an interprofessional relationship grid or blueprint can be designed that is task specific. Table 12.2 outlines an example of a task involving discharge planning of a stroke patient with five professions. The individual professionals have an understanding of their own roles, responsibilities, and relationships, but this may not be known to the other professions.

This blueprint grid presented in Table 12.2 illustrates that for the team to function the roles, responsibilities, and relationships of the individual professions must be understood by other team member. The IPE activity should therefore determine which profession will be the most appropriate for each of the roles, responsibilities, and relationships. The team can then progress from an individual professional identity to a team-based competency identity to achieve completion of the task and then will have developed a performance identity.

12.2.6.6 Entrustable Professional Activities

It has been argued that competencies can no longer be assessed in isolation (ten Cate 2005; Mulder et al. 2010; Sterkenburg et al. 2010). A complex task reflects complex integration of competencies needed in actual day-to-day practice for teams and; therefore, addresses authentic practice. An entrustable professional activity (see Footnote 2) designate for a team might be defined as:

> A team demonstrates the necessary attitudes, knowledge, skills and behaviors (competencies) to be trusted to independently perform this activity or task.

Table 12.2 Assessment identity blueprint

CASE: stroke rehabilitation discharge									
Composition	Structure			Function			Outcome		
	Professional identity			Competency identity			Performance identity		
	Role	Responsibility	Relationship	Role	Responsibility	Relationship	Role	Responsibility	Relationship
Physician									
Nurse									
Occupational therapist									
Physical therapist									
Speech-language pathologist									

In the development of an entrustable professional activity for IPE, five stages of decreasing supervision could be applied to teams as they progress from structure to function to task:

1. No task execution (occasional team observation)
2. Task execution under direct supervision (supervisor present)
3. Task execution with supervision quickly available (but requires regular debriefing)
4. Unsupervised practice (debriefing as required)
5. Supervision may be provided to other teams (teach is to learn twice).

Figure 12.4 presents how elements of structure, function, and outcome can be linked to professional profile, competence, and performance by using entrustable professional activities.

Using this approach in IPE could allow teams to map their performance to competency frameworks such as those developed by the Canadian Interprofessional Health Collaborative (2010) or Interprofessional Education Collaborative Expert Panel (2011). In addition, a performance framework could be used in curriculum planning and assessment for all professions using formative, and/or summative assessments. More importantly this can generate evidence regarding the collaborative performance levels in a program for—useful for both external and internal evaluation purposes.

12.2.6.7 The Use of Milestones

Interprofessional teams based in different clinical settings are composed in many different formulations. For example, there can be experienced teams with new members, or conversely, newly forms teams with members who have vast experience of teamwork. Whatever the team composition is in relation to different professions, they pass through the processes of forming, storming, norming, and performing (Tuckman 1965). Inevitably, different interprofessional teams will evolve from novice to expert teams at different rates. The temporal nature of team

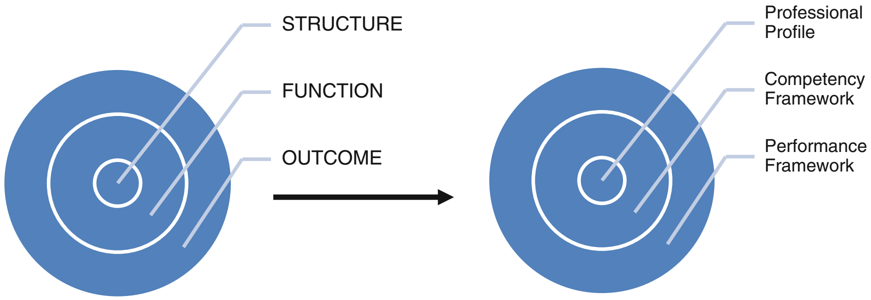


Fig. 12.4 Entrusted professional activities

Table 12.3 Milestone map for IPE competences

Competency	Stage (milestone) 1	Stage (milestone) 2	Stage (milestone) 3
Communication	Occasional use of appropriate communication	Effective use of appropriate communication	Excellent use of appropriate communication
Collaboration	Occasional team interaction	Effective use of appropriate team interaction	Excellent use of appropriate team interaction
Professionalism	Occasional use of appropriate attitudes ethics and behaviors	Effective use of appropriate attitudes ethics and behaviors	Excellent use of appropriate attitudes ethics and behaviors

development therefore needs to be taken into account when assessing the elements outlined above. A useful approach to measuring this developmental process (and different levels of attainment—from novice, to competent, to expert) has been described as milestones (ten Cate 2005).

Milestones are the abilities expected of a team at a defined stage of development, and as such, they should be based on a competency-based education approach (ten Cate 2013). Many current IPE competency frameworks only provide a definition for the completion of training in IPE (e.g., CIHC 2010; IPEC 2011). Milestones can mark achievable standards at different stages throughout IPE programs (Wagner and Reeves 2015). They can therefore help teams and educators identify the path to achieve the required competencies, adjust learning to team needs and abilities and also identify teams that may need additional training.

Using the competency domains of communication, collaboration, and professionalism (e.g., CIHC 2010), Table 12.3 outlines how a milestone map might appear. These domains are very broad and are often divided into subsets. For example teams progress along milestones by learning to establish rapport with patients and families, elicit and address patient's goals.

Table 12.3 can also provide an example of a blueprint for evolution from novice to expert. The stages will be determined by the competency framework used. A key point for team development is that teams may progress through stages based on competence at different rates. For example, a team may reach Stage 3 for communication but still be at Stage 1 for professionalism.

12.3 Conclusion

The assessment of IPE is fraught with complexity due to its application to individuals as well as interprofessional teams/groups. Although knowledge-based IPE assessments are a useful way of assessing individual understanding of IPE, there are limited as they fail to assess possible changes in behavior (performance). This chapter suggests using a multifactor approach to assessment by examining team structure (made up of individuals), the functions of the team (understanding their roles, responsibilities, and relationships) and outcomes (task completion). Such an approach allows team developmental milestones to be reached that can be assessed using entrusted professional activities.

Issues/Questions for Reflection

- In designing an IPE assessment have you clearly articulated its purpose? Are you clear what you will assess and how the assessment undertaken?
- What contents and outcomes will make up your assessment blueprint?
- What competency domains might you employ to assess collaborative performance?
- What elements will be in your IPE assessment matrix?
- Will you employ entrustable professional activities and milestones in your IPE assessment? If so how?

References

Barr, H., Koppel, I., Reeves, S., Hammick, M., & Freeth, D. (2005). *Effective interprofessional education: Assumption*. Blackwell, Oxford: Argument and Evidence.

Benner, P. (1982). From novice to expert. *American Journal of Nursing, 82*(3), 402–407.

Centre for the Advancement of Interprofessional Education (CAIPE). (2002). *Interprofessional education: A definition*. Available at: www.caipe.org/ipe-definition

Canadian Interprofessional Health Collaborative. (2010). *A national interprofessional competency framework*. Available at: http://www.cihc.ca/files/CIHC_IPCompetencies_Feb1210.pdf

Chumley, H. S. (2008). What does an OSCE checklist measure? *Family Medicine, 40*, 589–591.

Curran, V., Hollett, A., Casimiro, L., Mccarthy, P., Banfield, V., Hall, P., et al. (2011). Development and validation of the interprofessional collaborator assessment rubric (ICAR). *Journal of Interprofessional Care, 25*(5), 339–344.

D'Eon, M. (2004). A blueprint for interprofessional learning. *Medical Teacher, 26*(7), 604–609.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. New York, NY: Harper & Row.

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., et al. (2010). Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. *Lancet, 376*(9756), 1923–1958.

Hamdy, H. (2006). Blueprinting for assessment of health care professionals. *Clinical Teacher, 3*, 175–179.

Hammick, M., Freeth, D., Koppel, I., Reeves, S., & Barr, H. (2007). A best evidence systematic review of interprofessional education. *Medical Teacher, 29*, 735–751.

Institute of Medicine. (2013). *Interprofessional education for collaboration: Learning how to improve health from interprofessional models across the continuum of education to practice—Workshop summary*. Washington DC: National Academies of Practice.

Institute of Medicine. (2014). *Measuring the impact of interprofessional education (IPE) on collaborative practice and patient outcomes: A consensus study*. Washington DC: National Academies of Practice.

Interprofessional Education Collaborative Expert Panel. (2011). *Core competencies for interprofessional collaborative practice: Report of an expert panel*. Washington, DC: Interprofessional Education Collaborative.

Lidskog, M., Löfmark, A., & Ahlström, G. (2009). Learning through participating on an interprofessional training ward. *Journal of Interprofessional Care, 23*, 486–497.

Lundon, K., Kennedy, C., Rozmovits, L., Sinclair, L., Shupak, R., Warmington, K., et al. (2013). Evaluation of perceived collaborative behaviour amongst stakeholders and clinicians of a continuing education programme in arthritis care. *Journal of Interprofessional Care, 27*(401–407), 11.

Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, S63–S67.

Morison, S., & Stewart, M. (2005). Developing interprofessional assessment. *Learning in Health and Social Care*, *4*(4), 192–202.

Mulder, M., Gulikers, J., Biemans, H. J. A., & Wesselink, R. (2010).The new competence concept in higher education: error or enrichment? In D. Münk & A. Schelten (Eds.), *Kompetenzermittlung für die Berufsbildung. Verfahren, Probleme und Perspektiven im nationalen, europäischen und internationalen Raum* (pp. 189–204). Bundesinstitut für Berufsbildung, Bonn.

Pollard, K. C., & Miers, M. E. (2008). From students to professionals: Results of a longitudinal study of attitudes to pre-qualifying collaborative learning and working in health and social care in the United Kingdom. *Journal of Interprofessional Care*, *22*, 399–416.

Reeves, S. (2012). The rise and rise of interprofessional competence. *Journal of Interprofessional Care, 26*, 253–255.

Reeves, S., Lewin, S., Espin, S., & Zwarenstein, M. (2010). *Interprofessional teamwork for health and social care*. London: Blackwell-Wiley.

Reeves, S., Perrier, L., Goldman, J., Freeth, D., Zwarenstein, M. (2013). Interprofessional education: Effects on professional practice and healthcare outcomes (update). *Cochrane Database of Systematic Reviews 2013*, Issue 3. Art. No.: CD002213. doi: 10.1002/14651858.CD002213.pub3

Royal College of Physicians and Surgeons of Canada. (2005). *CanMEDS physician competencies*. Ottawa: RCPSC.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco, CA: Jossey-Bass Publishers.

Singleton, S., Smith, F., Harris, T., Ross-Harper, R., & Hilton, S. (1999). An evaluation of the team objective structured clinical examination (TOSCE). *Medical Education, 33*, 34–41.

Simmons, B., Egan-Lee, E., Wagner, S., Esdaile, M., Baker, L., & Reeves, S. (2011). Assessment of interprofessional learning: the design an interprofessional objective structured examination approach. *Journal of Interprofessional Care, 25*, 73–74.

Simmons, B., Egan-Lee, E., Wagner, S., Esdaile, M., Baker, L., & Reeves, S. (2012). *Interprofessional objective structured examination (iOSCE): Final evaluation report*. Toronto: University of Toronto.

Sterkenburg, A., Barach, P., Kalkman, C., Gielen, M., & ten Cate, O. (2010). When do supervising physicians decide to entrust residents with unsupervised tasks? *Academic Medicine, 85*(9), 1408–1417.

Symonds, I., Cullen, L., & Fraser, D. (2003). Evaluation of a formative interprofessional team objective structured clinical examination (ITOSCE): A method of shared learning in maternity education. *Medical Teacher, 25*(1), 38–41.

ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39*, 1176–1177.

ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of graduate medical education, 5*(1), 157–158.

Tuckman, B. (1965). Developmental sequence in small groups. *Psychological Bulletin, 63*, 384–399.

van der Vleuten, C. (2008). *Assessment past present and future, theories and concepts*. Keynote Presentation, Wilson Centre Research Day, University of Toronto, November 2008.

Wagner, S., & Reeves, S. (2015). Milestones and entrustable professional activities: The key to practically translating competencies? *Journal of Interprofessional Care, 29*, 507–508.

World Health Organization. (2010). *Framework for action on interprofessional education and collaborative practice*. Geneva: WHO.

Zwarenstein, M., Atkins, J., Hammick, M., Barr, H., Koppel, I., & Reeves, S. (1999). Interprofessional education and systematic review: A new initiative in evaluation. *Journal of Interprofessional Care, 13*, 417–424.

Chapter 13
Incorporating Groupwork into Performance Assessments: Psychometric Issues

Noreen M. Webb

Abstract Organizations increasingly rely on groups or teams to carry out many types of tasks. To assess individuals' capabilities in working with others and the performance or productivity of groups, groupwork must be incorporated into performance assessments. The purpose of this chapter is to enumerate and describe the challenges that emerge in performance assessments that include groupwork, and to offer suggestions for addressing the challenges and identify areas for future research. One set of challenges is that groups may function differently and in ways that do not align well with the goal of the assessment. Group dynamics influencing assessment scores include the nature of task-related interaction with others, lack of involvement of one or more team members, uncoordinated group communication, social-emotional processes, and division of labor. Another set of challenges is the large number of sources of measurement error that are unique to groupwork or that operate in new and different ways in groupwork settings. These sources, which have implications for validity and reliability, include variation due to group composition, role assignment in the group, task and type of task, occasion of observation, rater and type of rating or rating scale, and automatic coding and scoring. This chapter describes strategies for investigating and addressing the complexity involved in assessing groupwork performance, and describes implications for practice.

Takeaways

- Assessment of groupwork presents measurement challenges beyond those involved in individual-level assessment.
- Multiple factors influence how groups function and, consequently, the scores produced from an assessment, including group composition, roles in the group, type of task, occasion of measurement, type of rater, and rating scale.

N.M. Webb (✉)
Faculty of Education, UCLA Graduate School of Education and Information Studies,
Los Angeles, CA, USA
e-mail: webb@ucla.edu

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_13

- Obtaining reliable scores may require assessments to include multiple group compositions, group member roles, tasks, and raters.
- The optimal design of assessments with groupwork may differ according whether the objective is to measure the contributions or performance of individual group members or the productivity or performance of the group as a whole.

In the workforce, organizations increasingly rely on groups or teams to carry out many types of tasks (e.g., production, service, management, project, action and performing, advisory, negotiation, problem solving) in a wide variety of settings (e.g., military, government, civilian, and healthcare; Wildman et al. 2012). Large or complex tasks require teams to complete them successfully (e.g., designing new products); other tasks, by definition, involve multiple participants (e.g., negotiating agreements).

Employers view the ability to collaborate with others as a core twenty-first century competency that is more important than even English language ability or subject matter knowledge for both landing and keeping a job (National Research Council 2011). Leaders in healthcare professions increasingly recognize that outcomes such as patient safety depend on interprofessional collaboration and effective team communication. These views echo earlier reports of workplace know-how, such as the Secretary's Commission on Achieving Necessary Skills (SCANS 1999), which identified interpersonal skills—notably working in teams—as an essential competency of effective workers. The importance of these skills will only increase in the future (Griffin et al. 2012). Thus the primary reason for the pressure to include groupwork in performance assessments stems from the wish to assess individuals' capabilities in working with others to accomplish a common task. Additional reasons include assessing abilities that cannot be easily measured, or cannot be measured at all, in individual assessments, for equity concerns, and to send signals about desirable instructional practices.

The purpose of this chapter is to enumerate and describe the challenges that emerge in performance assessments that include groupwork, and offer suggestions for addressing the challenges and identify areas for future research. This chapter focuses on performance assessments in which individuals work together in small groups of size two or more to achieve a common goal, whether it is solving a problem, completing a task or activity, or producing a product. The conception of groupwork in this chapter is quite broad. Interactions among members of a group may take place face-to-face or virtually, and may occur synchronously or asynchronously. Group members may or may not play specific roles or be assigned specific activities, and groups may or may not follow scripts or instructions about how they should interact. Moreover, not all members of the group need be human, some may be computer agents playing the role of fellow examinees. This chapter also considers situations in which a human or computer confederate plays the role of non-peer (e.g., a trained interviewer or computer agent playing the role of an expert or native-speaking conversational partner in tests of oral language proficiency; a trained actor playing the role of a patient in tests of clinical skills in medicine).

The object of the measurement may be the group or individual members of the group. Moreover, the measurement may focus on processes occurring during the group's work, the product generated by the group, or individual or group performance exhibited during or after groupwork.

It should be noted that this chapter focuses largely on performance tasks in standardized assessment contexts, such as large-scale testing programs, certification or licensing efforts, or job-performance testing for industry. Less attention is paid to assessment tasks that might be embedded in instruction, or that may take place at the end of instruction, or to end-of-course evaluations to gauge individuals' performance or improvement. Nonetheless, the issues explored in this chapter are relevant to assessments carried out in instructional contexts as well, even though the requirements for reliability and validity may not be as high as in high-stakes or large-scale testing.

13.1 Why Might We Incorporate Groupwork into Performance Assessments?

Over the past several decades, educators, researchers, and policy makers have become increasingly interested in using performance tasks in large-scale assessment programs to assess what individuals know and can do. In contrast to conventional multiple-choice tests, performance assessments require examinees to respond to complex tasks that represent authentic or "real world" problem solving or performance situations and, ideally, do a better job in assessing examinees' higher order thinking skills, deep understanding of content, complex problem solving, and communication. While many performance assessments require only individual performance, they can be extended to involve individuals working in groups on a common task or set of tasks to reflect the value that educators, policy makers, the business community, and the general public place on the ability to communicate and collaborate effectively with others (Baron 1992). In addition to the usual challenges of measuring examinees' performance on complex tasks carried out individually, groups working on common tasks present unique or more complicated measurement challenges.

Performance assessments might include groupwork for a variety of reasons in order to:

- Measure collaboration and teamwork skills
- Measure group productivity or performance
- Measure individuals' ability to communicate about subject matter or job tasks
- Equalize intellectual resources among examinees
- Measure how well individuals can perform after groupwork practice

- Influence classroom instructional practices
- Increase the fidelity of the assessment to instructional or work activities

What are the collaboration and teamwork skills that figure so prominently in successful group performance and high-quality group productivity? Common themes running through the many taxonomies and frameworks of teamwork and teamwork skills (more than 138 models according to Salas et al. 2005) include adaptability (recognizing problems and responding appropriately), communication (exchanging clear and accurate information), coordination (organizing team activities to complete a task on time), decision-making (using available information to make decisions), interpersonal (interacting cooperatively with other team members), and leadership (providing structure and direction for the team, Chung et al. 1999; SCANS 1999). Similar core teamwork skills underlie healthcare team training programs, especially monitoring situations to make sure that work is proceeding as expected, providing mutual support to team members, directing and coordinating team members' actions, and initiating, acknowledging, and verifying communications (Baker et al. 2010). In the educational arena, subject matter competence includes being able to communicate one's ideas as well as respond to others' ideas (e.g., constructing viable mathematical arguments and critiquing the reasoning of others; National Governors Association for Best Practices, and Council of Chief State School Officers 2010). By definition, these collaboration, communication, and teamwork skills involve interaction with others. Incorporating groupwork into assessments provides a direct way of measuring these skills as well as the productivity and performance of the group as a whole.

Performance assessments might also include groupwork as preparation for subsequent individual performance. A groupwork phase that precedes a purely individual phase may be used to help equalize intellectual resources for individuals who have had less opportunity than others to learn the material. That is, the opportunity to share knowledge and understanding during groupwork practice may help level the playing field (Baron 1994; Neuberger 1993). Including preparatory groupwork prior to individual work may also make it possible to measure how well individuals can learn from working with others, which is consistent with a perspective of competence that sees learning as constructed in collaboration with others (e.g., Vygotsky 1978). Providing opportunities for groupwork and then assessing individuals after groupwork practice, then, may be seen as a way both to increase fairness and to measure how well individuals can learn from collaborative experiences.

Including groupwork on assessments also constitutes a signaling function for classroom instruction (Linn 1993; Stecher 2010). Decades of research in instructional settings show the power of collaborative groupwork for learning and other outcomes such as the development of pro-social attitudes (Webb and Palincsar 1996). Incorporating groupwork into assessments can serve as a policy lever to influence classroom instructional practices, especially if the assessments simulate a beneficial learning environment that affords high-quality collaboration.

Finally, groupwork on assessments can increase the fidelity of the assessment to instructional activities (and thus increase its instructional sensitivity or instructional validity, Popham 2007). Instructional activities incorporate groupwork to provide opportunities for collaboration, to afford opportunities to tackle large and complex tasks that individual students cannot easily complete, or both. For example, including partner work on science laboratory tasks in the classroom (e.g., investigating why an ice cube sitting on an aluminum block melts faster than an ice cube on a plastic block) provides opportunities for students to engage in collaborative scientific argumentation that fosters scientific literacy (e.g., discussions about gathering and making sense of data, generating and testing hypotheses, justifying explanations, critiquing viewpoints, Sampson and Clark 2011). Group preparation of multifaceted research reports on complex topics, such as the phototropic behavior of plants, is another example (Cohen and Cohen 1991). Assessment tasks that provide collaboration opportunities represent such instructional activities better than purely individual tasks.

13.2 Example Educational Assessments with Groupwork

A number of large-scale assessments have incorporated groupwork (or will do so) to serve one or more of the purposes described above. For example, in the 1990s, several state assessments incorporated collaborative groupwork alongside individual assessments in response to recommendations of state and national assessment standards. One was the three-part science task in the Connecticut Common Core of Learning Alternative Assessment Program, in which students first individually provided information about their science knowledge; then worked in three- or four-person teams to design, carry out, interpret, and summarize an experiment; and finally individually reflected about the group activity, independently analyzed and critiqued another group's report, and applied knowledge gained during the groupwork phase (Baron 1992, 1994). Students' performance was scored on all three parts. Other assessments with collaborative groupwork included the Connecticut Academic Performance Test (CAPT, Connecticut State Board of Education, 1996; Wise and Behuniak 1993), the Maryland School Performance Assessment Program (MSPAP, Maryland State Department of Education 1994), the California Learning Assessment System (CLAS, Saner et al. 1994), and the 1994 Kansas Science Assessment (Pomplun 1996).

Currently, the Smarter Balanced Assessment Consortium uses groupwork in some assessment tasks. For example, in a sample Grade 11 Performance Task on Thermometer Crickets, students work in small groups during classroom instruction to build background knowledge (e.g., why crickets chirp primarily at night), engage in a class discussion (e.g., interpretation of data about cricket chirping in different conditions), and then complete tasks individually (e.g., organizing and analyzing data about the relationship between temperature and crickets' chirping rates, http://www.smarterbalanced.org). The scoring rubric focuses on student performance on the individual task that follows the class discussion (e.g., plotting the data points, modeling and interpreting the relationship). Groupwork appears internationally in high-stakes assessments as well, such as the Singapore A-levels required for university admission (National Research Council 2011).

In the near future (planned for 2015), another international assessment– the Programme for International Student Assessment's (PISA) test of individuals' collaborative problem solving skills–will have examinees interact with one or more computer-simulated collaborators to solve problems such as finding the optimal conditions for fish living in an aquarium, or producing an award-winning logo for a sporting event (Organisation for Economic Co-operation and Development, March 2013). Examinees will be scored on their behavior when interacting with the computer agent (e.g., communicating about the actions to be performed to complete the task) and on their responses to multiple-choice and constructed-response probes placed within the unit (e.g., write an email explaining whether there is group consensus on what to do next). The scoring will focus on three competencies: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization. Specific skills within these dimensions include, for example, coordination, explanation, filling roles, argumentation, and mutual regulation. Along similar lines, the National Center for Education Statistics within the U.S. Department of Education is also planning how to assess collaborative problem solving in the National Assessment of Educational Progress (2014 NAEP Innovations Symposium: Collaborative Problem Solving held in Washington, DC, September 29, 2014).

13.3 Concerns About Construct Interpretations: The Effect of Group Processes

Groupwork introduces complexities not found in purely individual assessments. Regardless of the reason(s) for incorporating groupwork into an assessment, the mere presence of groupwork renders invalid interpretation of scores as reflecting unassisted individual competence. For example, in the 90-minute Connecticut Academic Performance Test in language arts, insertion of a brief 10-minute group discussion partway through the test improved students' understanding of the story and their scores on the test (Fall et al. 2000). Similarly, designing and carrying out science experiments and investigations in pairs on the California Learning Assessment System helped students develop new ideas, knowledge, and understanding (Saner et al. 1994). Students' scores on the tests, then, reflected a combination of their own competence and what they learned or gained from the groupwork experience.

Further complicating the interpretation of scores from assessments with groupwork, some groups function more effectively than others for reasons that may or may not align well with the construct(s) of interest. Several ways in which group functioning may differ, with consequences for assessment scores, include:

- Task-related interaction with others
- Lack of involvement
- Uncoordinated group communication
- Social-emotional processes
- Division of labor

Whether these processes are beneficial or detrimental for performance scores will depend on the target construct.

13.3.1 Task-Related Interaction with Others

Group members can interact with each other around the task in a great many ways, such as sharing information and ideas; building on each other's ideas to solve a problem, constructing new knowledge, or completing a task; engaging in conflicts and resolving disagreements; and seeking, giving, and receiving help (Webb and Palincsar 1996). Giving and receiving help, for example, can promote learning, and thus improve individuals' scores after groupwork practice, by encouraging individuals to rehearse information, reorganize and clarify material in their own minds, recognize misconceptions and gaps in understanding, strengthen connections between new and previously learned information, and develop new perspectives. Engaging in helping behavior may also lead raters to score examinees highly on communications skills. In terms of group productivity, however, spending time to ensure that everyone understands the material may slow the group down and prevent it from solving the problem or completing the task. In that case, suppressing the participation of less capable members or individuals who are experiencing difficulty may help the group improve its performance score.

13.3.2 Lack of Involvement

Not all members of a group may contribute to groupwork. Consider, for example, social loafing, or diffusion of responsibility, which arises when one or more group members sit back and let others do the work (Karau and Williams 1993; Slavin 1990). Individuals may go along for a "free ride" if they believe that their efforts cannot or will not be identified or are dispensable (Kerr and Bruun 1983; Levine and Moreland 1990). In addition, members of a group may not participate if they are discouraged, unmotivated, unrecognized, intellectually snobbish, intentionally passive, or involved in something else (Mulryan 1992). Whether and how diffusion of responsibility influences scores on an assessment depend on the focus of the measurement: individual contributions or group productivity. Uninvolved group members will likely receive low individual scores on their subject matter communication skills and contributions to teamwork, and possibly also individual scores after groupwork practice, especially if they lacked relevant knowledge or skills coming into the assessment (Webb 1993). How social loafing might affect *group* scores depends on group members' capabilities and task attributes. On the one hand, social loafing may be detrimental for group productivity if the social loafers have necessary skills for the group to accomplish the task (which may be especially relevant for nonroutine tasks that do not have well-specified procedures, Cohen and Cohen 1991), or if social loafing becomes contagious (Salomon and Globerson 1989). On the other hand, groups may function better and complete tasks more effectively if some members keep quiet, especially if they do not have new or productive ideas to contribute.

13.3.3 Uncoordinated Group Communication

Instructional research shows that opportunities for groups to benefit from information sharing may be lost when group members do not coordinate their communication. In uncoordinated conversations, individuals advocate and repeat their own positions and ideas, ignore others' suggestions, reject others' proposals without elaboration or justification, and interrupt others or talk over them (Barron 2000). In highly coordinated groups, in contrast, members acknowledge and elaborate upon each other's ideas. Although lack of coordination of group members' efforts on assessments with groupwork can impede group functioning and reduce the quality of the group's product (and thus their group productivity score), group members who actively promote their own ideas (even if they do not engage with others) may nonetheless receive high individual communication scores.

13.3.4 Social-Emotional Processes

Negative social-emotional processes, such as being rude, hostile, unresponsive, and domineering, can impede group functioning in multiple ways, such as causing group members to withhold correct information from each other and to reject viable suggestions posed by others (Chiu and Khoo 2003). While such processes can negatively impact group productivity and reduce opportunities for group members to benefit from groupwork practice (Webb and Mastergeorge 2003), they may not be detrimental for individuals' communication scores (e.g., dominant group members being marked high for their frequent contributions). Positive social-emotional processes such as cooperativeness, cohesiveness, team spirit, and liking of team members may improve group productivity and performance unless the good feelings arise out of suppression of disagreements, which can lead to reduced group productivity and opportunities to benefit from groupwork practice (Webb and Palincsar 1996).

13.3.5 Division of Labor

Division of labor, that is, dividing the task into parts and assigning different group members responsibility for completing different parts, may be a productive, efficient, or even necessary, strategy for accomplishing group tasks (Salas et al. 2000) and may, consequently, increase group performance or group productivity scores. However, if this strategy curtails interaction among group members, it may produce underestimates of individuals' scores on, for example, their ability to collaborate with others, communicate about the subject matter, and apply or synthesize knowledge gained during groupwork practice.

In summary, the nature of group processes that arise in a particular groupwork session may greatly impact scores of groups and/or their members. Some influences may be construct-relevant (a group with highly coordinated communication receives a high score on teamwork skills), while other influences may be construct irrelevant (an individual receives a low communication score because the group divided up the labor to accomplish the task and spent little time discussing it). The following sections consider influences on processes and performance in the groupwork setting that may cause scores to vary and, consequently, impact validity and/or reliability of score interpretations.

13.4 Sources (Facets) of Measurement Error Specific to Groupwork

Much has been written about important sources of unwanted variation in individual-level assessment scores, such as variability due to the sampling of tasks, occasions, raters, rater types, and type of rating or rating scale (e.g., Lane and Stone 2006). These sources (facets) of measurement error figure prominently in assessments with groupwork as well. In addition, the groupwork setting introduces new sources of score variability that do not pertain to individual assessments, such as the composition of the group and the roles played by group members. This section addresses sources of measurement error that are unique to groupwork, as well as ways in which sources of error relevant for individual assessment may operate in new or different ways in groupwork settings. The sources of variability described below are relevant for both group scores and scores of individuals within groups.

13.4.1 Variation Due to Group Composition

The composition of the group can vary along a great many dimensions including group member knowledge and skill level, gender, personality, motivation, acquaintanceship, status, popularity, attractiveness, ethnic background, race, and other demographic characteristics. The large instructional literature on cooperative or collaborative learning in the classroom shows a marked influence of group composition on many outcomes, including group processes, group performance, and student learning. Similarly, research in organizations (e.g., industry, military) shows that team composition on, for example, cognitive and psychomotor abilities, organizing skills, cooperativeness, team orientation, greatly influences team functioning and success (National Research Council 2013; Mathieu et al. 2014). Emerging evidence shows that group composition, especially the homogeneity of the group in terms of, for example, achievement level or perceptions about the task or teamwork skills, matters in assessment situations, too. Some studies, for example, have found homogeneous groups to produce higher scores than heterogeneous groups on collaborativeness, resolution of cognitive conflicts, and communication when engaging in complex mathematics tasks (Fuchs et al. 1998) and aircraft simulation tasks (Mathieu et al. 2000). Other studies, in contrast, have found that diversity of multiprofessional healthcare teams is positively related to team processes such as the extent to which the group reflects on its processes and strategies, how safe group members feel to express minority views, and how frequently group members interact with each other (Fay et al. 2006).

Of particular interest is the *combined* influence of group composition and group members' own characteristics on group processes and outcomes, such as average-ability students, low-status students, and girls being less active and learning less in heterogeneous than in homogeneous classroom groups (Webb and Palincsar 1996). This combined influence appears in assessment studies as well, such as high-ability students showing higher scores in homogeneous groups and in high-functioning heterogeneous groups than in poorly functioning heterogeneous groups (e.g., group members failing to answer each other's questions, failing to explain their answers, insulting others, Webb et al. 2002).

Growing concern about the combined effects of test taker characteristics and group composition appears in recent research on testing the language capability of students learning English as a second language. Increasingly, group oral tests are being introduced into low-stakes and high-stakes tests in order to assess communication ability in a more naturalistic setting than, say, an oral interview with an examiner. Characteristics such as gender, personality (especially introversion-extraversion), language proficiency, and acquaintanceship (friends, acquaintances, and strangers) have different effects on communication depending on how test takers are paired (e.g., Ockey et al. 2013), as well as on the size of the group (e.g., introverted students participate more actively in smaller groups than in larger groups, Nakatsuhara 2011).

Recent evidence suggests that behavior of, and scores assigned to, the same individual may change from one group composition to another. One example is a study conducted in a managerial assessment center, commonly used in the business community to gauge prospective employees' job skills such as communication, consideration/awareness of others, drive, influencing others, organization and planning, problem solving, leadership, learning from others, fostering relationships, and managing conflict (Collins and Hartog 2011). Assessment center exercises include, for example, role plays in which an examinee presents a business plan or carries out a coaching conversation with a trained role player who responds in prescribed ways to the test taker's actions. Hoffman and Meade (2012) observed managers in an executive MBA program in two such role play exercises. In one, test takers interacted with a high-performing but interpersonally abrasive role player; in the other, they interacted with an average performing (and presumably nonabrasive) role player. Ratings of group process dimensions (e.g., oral communication, sensitivity, leadership, and confrontation) for the two role play exercises did not correlate highly, showing that scores may not be generalizable from one group composition to another, and that test takers may need to be observed in a large number of group compositions for generalizable results.

Given the difficulty of ensuring that group compositions are similar across groups, considerable interest lies in controlling variation in group composition by standardizing attributes and behavior of group members. Hoffman and Meade's (2012) study points to one method of standardization: using scripted confederates to play the role of group partners. Scripted group members have been used in a variety of assessment situations, such as live actors playing the role of patients in medical performance tasks involving history taking, physical examination, and patient education (e.g., Richter Lagha et al. 2012), confederates posing as team members (co-pilots) and following a script that presents prescribed conflict situations for test takers to resolve in flight simulation tasks, such as making blatant errors and remaining indecisive regarding critical decisions (Smith-Jentsch et al. 1996), and trained interviewers engaging in structured conversations with test takers in tests of oral proficiency (e.g., the American Council on the Teaching of Foreign Languages Oral Proficiency Interview; See http://www.actfl.org/professional-development/certified-proficiency-testing-program/testing-proficiency).

Unless the behavior of confederates is highly controlled, however, their attributes and behaviors can introduce error into the measurement. Lazaraton (1996), for example, documented multiple ways in which interviewers in oral proficiency assessments might influence conversations (and, hence, test takers' language scores), such as interviewers completing test takers' sentences or thoughts, echoing or correcting test-taker responses, repeating questions using slowed or over-articulated speech, or rephrasing questions. On the other hand, tightly scripting confederates to eliminate such influences may produce unnatural conversational discourse (Malone 2003).

As an alternative to using human confederates, test takers might interact with computer conversational agents (simulated group members) that are programmed to respond to test taker behavior in certain ways. The use of computer agents is at the heart of the 2015 PISA effort to measure collaborative problem solving competencies (Organisation for Economic Co-operation and Development, Organisation for Economic Co-operation and Development, March 2013). Conversational agents will represent peers with a range of skills and abilities and other characteristics, as well as behavior (team members who initiate ideas and support and praise others versus team members who interrupt and criticize others and propose misleading strategies). Pilot studies have found similar levels of motivation to accomplish the task, time on task, and problem solving success among students interacting with a computer agent and students working in the same online environment but with a human partner (Rosen and Tager 2013). Computer agents may also play the role of a non-peer. For example, in Alelo Inc.'s program to teach foreign languages and assess students developing proficiency in a new language, students interact with a native-language-speaking avatar to carry out tasks such as negotiating between conflicting parties to resolve an argument (Soland et al. 2013).

Whether computer agents can be designed that will reliably mimic realistic conversational partners is not known. For example, research on AutoTutor, a computer program that converses with students using natural language, shows that the computer does not reliably detect and classify emotions (e.g., frustration, confusion, and surprise), makes errors in interpreting the content of students' utterances (especially students' questions), and sometimes responds inappropriately due to misclassifying students' speech acts (Graesser et al. 2008). Limitations of a computer agent's communication facility may lead human participants to respond in unnatural ways, thus calling into question the validity of scores derived from human–computer interaction.

In conclusion, a number of questions remain to be answered, including the extent to which interacting with computer partners generalizes to interacting with human partners, whether using computer agents as partners produces groupwork experiences that are comparable from test taker to test taker, how many standardized group compositions are needed for generalizable scores, and how to select group compositions to represent the target domain of group compositions.

13.4.2 Variation Due to Role Assignment in Groupwork

Instructional research shows that role specialization can influence groupwork. To raise the level of discussion in groups, students can be assigned various roles such as recaller or learning leader roles to summarize the material and listener to ask questions, detect errors, and identify omissions in learning leaders' summaries (O'Donnell 1999). Assignment of roles may influence group process. Schellens et al. (2005) also found that assigning students the roles moderator, theoretician, summarizer, and source searcher in asynchronous online discussion groups produced more high-level communication about the task (e.g., testing and revising new ideas constructed by the group) than did group discussion without role assignment. Even in the absence of explicit role assignment, group members may assume specific roles that influence group dynamics, such as students positioning themselves as experts and novices and exhibiting asymmetrical teacher–learner interaction (Esmonde 2009).

Recognizing that role assignment may influence group collaboration, PISA plans to include tasks that differ according to role structure: Some tasks will have symmetrical roles (every group member has the same role) and others will have asymmetrical roles (different roles are assigned to different group members, such as scorekeeper versus machine controller, Organisation for Economic Co-operation and Development, March 2013).

13.4.3 Variation Due to Type of Task

There is increasing recognition, especially in research on managerial assessment centers, that the type of task may influence group processes and outcomes of groupwork. For example, role play exercises, simulated interviews, and leaderless group discussions are designed to call on different groupwork skills (Howard 2008) and may activate expression of underlying traits to different degrees and in different ways (such as extraverted test takers exhibiting influence more during leaderless group discussions than when giving oral presentations, Lievens et al. 2006). Indeed, reviews and meta-analyses of assessment center research show that ratings across different types of exercises are weakly to moderately correlated, even for the same groupwork dimension (e.g., Arthur et al. 2003; Bowler and Woehr 2006; Lance et al. 2010). Speer et al. (2013) explored this issue further by having experts evaluate the similarity of assessment center exercises and then examining correlations for more and less similar exercises. Although correlations between exercises for the same dimension rating (e.g., lead courageously, influence others, fostering teamwork, build relationships, manage disagreements, and fostering open communication) were fairly low overall, correlations between exercises judged to be more similar (e.g., role play and group discussion) were significantly higher than correlations between exercises judged to be less similar (e.g., role play and situational interview).

The type of groupwork task may also change the *distribution* of group members' contributions within the group. One dimension of task type is the degree of structure, such as tasks with well-defined procedures and answers that can be completed by one person (called disjunctive tasks, Steiner 1972) versus tasks with ill-structured solutions that cannot be completed very well by a single individual due to complexity or because no individual is likely to have all of the necessary expertise (Cohen 1994; Cohen and Cohen 1991). Chizhik et al. (2003) found that ill-structured tasks promoted more equally distributed participation among group members than did well-structured tasks.

Acknowledging that different types of tasks may require different groupwork skills, PISA plans to include different types of collaborative problem solving tasks that elicit different types of groupwork interactions and problem- solving behaviors. A possible typology of tasks includes "(a) group decision-making tasks (requiring argumentation, debate, negotiation, or consensus to arrive at a decision), (b) group coordination tasks (including collaborative work or jigsaw hidden profile paradigms where unique information must be shared), and (c) group production tasks (where a product must be created by a team, including designs for new products or written reports)" (Organisation for Economic Co-operation and Development 2013, p. 22). The variety of task types needed to represent the domain of task types well is not yet known.

13.4.4 Variation Due to Task

Consistent with task variability in individual performance assessment scores (e.g., Shavelson et al. 1993), group process and group performance may be quite variable even across tasks designed to be similar and to require similar competencies. Evidence of task variability comes from a variety of groupwork settings, such as military team simulations, simulation of medical-patient interactions, simulation of medical operating rooms, and simulation of teams in business and management. For example, Brannick et al. (1995) designed two simulated military air crew missions to be comparable in terms of group processes (e.g., assertiveness, decision-making, communication, and adaptability) and performance expectations (e.g., performance of the designated pilot). Despite the careful matching of tasks, correlations between scores on the two tasks were low on both ratings of group processes and performance. Similarly, in simulated operating room theaters, Morgan et al. (2007) found significant differences between common emergency obstetric scenarios in team functioning (e.g., information sharing, coordination among team members).

High-stakes assessments of medical students' clinical skills, such as an objective structured clinical examination (OSCE) in which examinees interact with standardized patients (specially trained actors) on multiple tasks representing common medical situations (sometimes called cases or stations), also show large variability in performance from task to task. For example, Guiton et al. (2004) found large variability across tasks for communication skills such as effectively gathering information, listening actively, and establishing professional rapport. Richter Lagha et al. (2012) found such large variability across tasks in scores based on history taking, physical examination, and patient education that they estimated that increasing the number of tasks on the exam from a manageable 6–8 to an unwieldy 24 would yield, at best, only moderate dependability of scores. It should be noted, however, that examinees encountered a different simulated patient for each task in these studies, so effects of the task and simulated patient were confounded. Disentangling these effects requires an alternative measurement design, such as having examinees complete multiple replicates of the same tasks, each with a different standardized patient, or training standardized patients on multiple tasks so that examinees encounter the same standardized patient on multiple tasks.

13.4.5 Variation Due to Occasion

Consistent with occasion variability in individual performance assessment scores, group process, and group performance may be quite variable across occasions. Some studies show improvement in groupwork scores over time (such as improvement in students' negotiation skills from one session to the next, O'Neil et al. 1992; Wimmers and Lee 2014), although the improvement may taper off over time (such as teamwork skill scores increasing over the first four fighter aircraft simulation missions and then remaining level after that, Mathieu et al. 2000). Other studies show instability in the relative standing of contributions of individual group members across occasions (such as Kenderski's 1983 finding that some students exhibited high levels of help-seeking behavior on one occasion, while other students did so on another occasion; see also Webb 1984). On the other hand, some evidence indicates that groupwork behavior may be more stable across time intervals within the *same* occasion. For example, for medical and psychology students working in dyads in videoconference settings to diagnose psychiatric cases, Meier et al. (2007) reported substantial consistency in scores on group process dimensions (e.g., sustained mutual understanding, dialogue management, and information pooling) over three time blocks in the same session. This result raises the possibility that it may not always be necessary to rate entire groupwork sessions to produce dependable group process scores.

13.4.6 Variation Due to Type of Rater

As is the case for individual assessments, assessments with groupwork can use expert observers (rating live or recorded groupwork), peers, or examinees themselves.[1] In the groupwork context, in contrast to individual assessments, peers are typically the other members of groups performing groupwork activities. Team members may rate themselves, each other, or their team as a whole on contributions to the team's work, interactions with teammates, contributions to keeping the team on track, expectations for quality, possession of relevant knowledge, skills, and abilities, listening ability, appreciation of different points of view, consensus-reaching skills, conflict-resolutions skills, ability to synthesize the team's ideas, and skills in involving others (Loughry et al. 2007; Taggar and Brown 2001; Zhang and Ohland 2009). As another example, members of triads of examinees in managerial assessment centers may evaluate their peers' level of activity, persuasiveness, and clarity of communication in decision-making tasks (Shore et al. 1992). Self and peer ratings also figure prominently in social network analysis of groups, such as the prevalence of adversarial relationships in the team (a group level score, Baldwin et al. 1997), or a group member's centrality in being asked for advice (an individual-level score, Sparrowe et al. 2001).

Although peer and self-ratings are less resource intensive than expert ratings based on observations of groupwork (Dyer 2004; Salas et al. 2003), lack of convergence with expert ratings is one reason why peer and self-ratings are unlikely to

[1] It should be noted that other self-rating methods for gauging teamwork skills include questionnaires asking respondents to rate their own skills (e.g., "I am a good listener"), and multiple-choice situational judgment tests asking examinees to pick the best option for resolving hypothetical groupwork scenarios or pick the option that best represents how they would react (National Research Council 2011; Wang et al. 2009). Because these measures typically are not directly tied to actual groupwork activities, they are not considered further here.

be used in high-stakes or summative assessments. Findings reported include (a) low to moderate correlations for cooperation, giving suggestions, accepting suggestions between expert observer and peer ratings of dyads flying simulated aircraft missions (Brannick et al. 1993), (b) low correlations between medical students' self-reports of their behavior when interacting with standardized patients in a clinical performance assessment and experts' ratings of videotapes of the same encounters (Richter Lagha 2013), and (c) self-ratings, peer ratings, and observer-ratings giving rise to different pictures of communication networks, and the centrality of specific individuals within them, in social network analyses (Bernard et al. 1979/1980; Kilduff et al. 2008; Kumbasar et al. 1994).

Another issue regarding peer- and self-ratings is their lack of independence. When team members rate each other, themselves, or their team as a whole on, say, contributions to the team's work, interactions with teammates, possession of relevant knowledge, listening ability, appreciation of different points of view, and conflict-resolutions skills (e.g., Loughry et al. 2007; Ohland et al. 2012; Taggar and Brown 2001), they are themselves participants in the groupwork experience they are being asked to rate. Nonetheless, members of the same group may not agree on well their team functioned (e.g., Morgan et al. 2007).

13.4.7 Variation Due to Type of Rating or Rating Scale

When rating the group as a whole or the behaviors of individual group members, multiple types of ratings or rating scales are available. Raters may code the presence or absence of specific events that are expected to take place at a particular point in time (e.g., providing information as required or when asked, asking for clarification of communication, verbalizing plans for procedures/maneuvers; Fowlkes et al. 1994), the frequency of group processes (e.g., making contributions to groupwork discussions that refer to other group members' ideas; Rimor et al. 2010; or offering justified claims during argumentation; Weinberger et al. 2010), or the quality of specific behaviors observed (e.g., the effectiveness of conflict resolution, Fuchs et al. 1998; the quality of mutual support team members give each other during a critical phase of an exercise, Macmillan et al. 2013). Or raters may score groups or group members on general process dimensions (e.g., the quality of information exchange, communication delivery, supporting behavior, initiative and leadership; Smith-Jentsch et al. 2008; sustained mutual understanding, dialogue management, information pooling, Meier et al. 2007; conflict resolution, collaborative problem solving, and communication, Taggar and Brown 2001).

The evidence about the convergence of scores from different types of ratings or rating scales is mixed, and is too limited to draw general conclusions. On the one hand, Macmillan et al. (2013) found substantial agreement between scores from analytic scoring of observable behaviors (e.g., rating of the team providing mutual support) and judges' overall ratings of team functioning on a scale from 1 to 5. On the other hand, Ohland et al. (2012, p. 625) reported modest correlations between scores on Likert-type and behaviorally anchored rating scales among peers who rated their teammates' contributions to the group's work. The first scale required team members to rate each other on specific items from five broad categories using Likert scales (strongly disagree to strongly agree). For example, items representing the category "contributing to the team's work" included "did a fair share of the team's work," "fulfilled responsibilities to the team", and "made important contributions to the team's final product." The second scale required team members to assign each other a single rating for broad categories each defined by a set of behaviors. For example, raters gave the highest rating for "contributing to the team's work" if they judged the teammate to behave in the following ways: "Does more or higher quality work than expected; makes important contributions that improve the team's work; helps to complete the work of teammates who are having difficulty."

13.4.8 Agreement Among Raters

As is the case for individual assessments, studies of observations of groupwork often report moderate to high agreement among raters, for example, rating live group interaction (e.g., the number of times 4th-grade students provide explanations to other students when solving mathematics problems, Roschelle et al. 2009; quality of support and problem solving suggestions offered to others in multidisciplinary teams in the medical operating theater, Mishra et al. 2009), rating videotaped groupwork (e.g., the frequency of informing other group members of critical information when flying simulated aircraft missions, Brannick et al. 1995), judging audiorecorded groupwork (e.g., information exchange, communication delivery, and initiative and leadership in Navy teams, Smith-Jentsch et al. 2008), rating online groupwork (e.g., the quality of argumentation in groups tasked with using theories about motivation and learning to understand and explain a student's poor performance in a mathematics course, Stegmann et al. 2012), and rating examinees' oral language proficiency during in-person or phone interviews (Ferrara 2008). Moreover, research shows that a feasible amount rater training (e.g., 18 h of practice rating and debriefing) can markedly reduce the magnitude of discrepancies between novice and expert raters' judgments of teamwork behavior (e.g., when rating the quality of communication, coordination, and co-operation in surgical teams, Russ et al. 2012).

Raters do not always agree even moderately, however, for reasons that may be specific to the groupwork setting. For example, raters evaluating conversations among pairs of examinees in a second language speaking test showed low agreement about examinees' language fluency and effectiveness (May 2009). Raters interpreted the same conversations very differently. For example in "asymmetric" interactions (one examinee more talkative than the other), one rater may have perceived the less dominant partner as loafing (and downgraded that examinee as a result) while another rater may have perceived the same examinee as being unfairly suppressed (and upgraded that examinee to compensate for the unfair pairing). Such results show that raters need training in how to take into account the possible influence of different patterns of interaction among examinees (e.g., asymmetric, parallel, and collaborative, Galaczi 2008) on their ratings of individuals' competencies such as oral language skills.

13.4.9 Automatic Coding and Scoring of Group Processes

Because human coding and scoring of group processes are very time-consuming and expensive, researchers are exploring automatic coding and scoring, especially in online environments that capture interaction automatically in data log files (e.g., capturing information flow in online forums to investigate networks of student participation, Zhang et al. 2009). Approaches for automatically scoring content of interaction include classifying transcripts of conversations (either between students, or between student and computer agent) according to their similarity to known text (Foltz and Martin 2008), and applying automatic speech recognition to analyze spoken communication (Johnson 2010). Researchers are investigating automatic scoring of groupwork interaction. Measures include (a) how frequently or well members support each other (providing backup, correcting errors, Dorsey et al. 2009) or how frequently students refer to each other's contributions, formulate counter-arguments, and collaboratively apply scientific concepts to solve problems (Rose et al. 2008); (b) identifying the roles that individuals play at any given point in time (e.g., directing groupwork, asking questions of others, encouraging participation, Goodman et al. 2005); (c) scoring speaking and understanding of a foreign language when interacting with an avatar (a realistic, computer-generated human being) in simulated face-to-face conversations (Johnson 2010); and (d) scoring students' communication about scientific methods (e.g., identifying flaws with news articles or blogs about science) when interacting with computer agents during a computer game (Forsyth et al. 2013).

Emerging evidence about the agreement between human judges and computer programs when coding the text of communications is mixed. Rose et al. (2008) reported fairly high agreement between human coders and text classification algorithms for some group processes such as connecting arguments to create a group argumentation pattern, although not for others (e.g., referring to the contributions of their group partners). The higher reliability indices compare favorably to agreement among human raters using similar coding schedules (e.g., Schoor and Bannert 2011). A major challenge for automatic scoring is how to construct classification algorithms that will transfer well between group discussion data for different topics and contexts (Mu et al. 2012).

An approach that bypasses the need to code ongoing interaction constrains the communication among group members to predefined instances of specific categories. Individuals choose from a menu of messages (derived from previous instances of unconstrained communication) that experts have judged to represent dimensions such as adaptability, communication, coordination (Chung et al. 1999), or building a shared understanding, problem solving, and monitoring progress (Rosen and Tager 2013). The number of times an individual sends messages in a particular category, such as decision-making, forms the basis for the individual's decision-making score. Menu-based interfaces may apply to fine-grained skills such as communicating with team members about the actions being performed, monitoring and repairing shared understanding, and prompting other team members to perform their tasks. For example, in the planned PISA assessment on collaborative problem-solving, test takers will be awarded points for taking specific actions such as asking the computer agent for the agent's point of view before implementing a plan (corresponding to the skill of building a shared representation) or monitoring whether the computer agent follows the plan as discussed (corresponding to monitoring and repairing the shared understanding, Organisation of Economic Co-operation and Development 2013). How well-constrained communication using predefined options maps onto natural communication without constraints remains to be investigated.

13.4.10 Variation Due to Random Coding Errors and Data Entry Mistakes

As is the case for individual assessments, unsystematic variation due to random coding errors and data entry mistakes can arise in assessments with groupwork. Here, however, a single error or unsystematic event can influence the scores of multiple test takers simultaneously and in different ways. For example, if one examinee's microphone malfunctions halfway through groupwork, reducing that examinee's contributions that are available for coding, the examinee may be scored too low while another examinee (whose contributions make up a larger share of the total as a result) may be scored too high. As another example, a rater who does not realize that two ideas are voiced by different individuals may credit one individual for both ideas, inflating the score for one individual and depressing the score for another. How effective usual strategies for minimizing the effects of errors on individual assessments, such as using multiple raters, multiple tasks, additional rater training, making imputations, or adjustments for missing data, will be for assessments with groupwork remains to be investigated.

Random errors may also affect group level scores and their reliability. Recent research in social network analysis suggests that the effects of unsystematic coding errors and data entry mistakes on the reliability of groupwork measures (such as an individual's centrality in a network) may depend on the particular kind of error in combination with both the particular set of relationships among team members and the particular role of the individual within the group. Wang et al. (2012) examined the effects of different types of measurement error in network data, such as the omission of a relationship between two members of a network, misspelling of a group member's name leading to the same individual being counted twice, or two individuals having the same name leading to them being considered to be the same person. They reported that some kinds of errors (mistakenly omitting a link between individuals) pose a bigger problem than other kinds of errors (mistakenly omitting an individual from the network), and the network measures for some types of networks (e.g., few clusters or subgroups) are more resistant to these errors than are other types of networks (e.g., many clusters or subgroups). Guided by their results, Wang et al. recommended targeted strategies for gathering additional data or cleaning the data to improve reliability, such as gathering additional data (or cleaning the data) for highly active individuals rather than for all individuals. This intriguing notion that it may be productive to collect additional observations for some individuals but not others, depending on their role in the network, may apply to groupwork scores other than network measures.

13.4.11 Implications for Validity and Reliability

All of the facets described above give rise to variability in assessment scores. An important question follows, namely, how to estimate the number and variety of conditions needed for dependable measurement of the target skills. Generalizability theory (Brennan 2001; Cronbach et al. 1972; Shavelson and Webb 1991), an important tool for understanding and estimating reliability and validity, can serve us well here. Generalizability studies estimate the magnitude of sources of error, and decision studies use that information to design a time- and cost-efficient measurement procedure.

Using the language of generalizability theory, the facets correspond to sources of error variation. The facets help define the possible *universe* of scores of interest. Specifically, the universe is defined by all combinations of conditions of the facets. Ultimately, we would like to know the universe score for an individual (e.g., an individual's ability to collaborate with others) or for a group (e.g., the quality of a group's solution to a complex problem) where the universe score is defined as the average of the individual's (or the group's) scores in the universe. For example, we want to know an individual's ability to collaborate with others across all possible group compositions in which the individual might work, all types of tasks, all occasions of test administration, all methods of coding and scoring, and so on. The question becomes, then: how well can we generalize from the observed score for an individual or for a group based on the particular instances of groupwork on an assessment to the universe score?

In considering how to answer this question, it is helpful to differentiate facets related to validity and those related to reliability. Facets that influence the meaning of the construct we term validity facets. If there is variation due to conditions of a validity facet, and the assessment includes some conditions but not others, the observed score may not represent the construct of interest. For example, we might be interested in test takers' teamwork skills whether they work with others during face-to-face interaction or in online interaction. If the two modes of communication generate substantially different scores, but only online communication is included in an assessment, generalization from the observed score to the universe score of interest will be severely compromised. As another example, if different types of rating systems (e.g., event-based ratings vs. ratings of general dimensions) do not converge, and only one rating type is used to rate observations of groupwork in an assessment, then the assessment scores may not generalize to the universe score of interest.

Efforts to standardize validity facets by purposively choosing conditions to include on the assessment (e.g., online interaction only) would produce scores that may generalize to a much more restricted universe than intended. Efforts to standardize validity facets may also change the nature of the generalization entirely. For example, if interacting with other people and interacting with a scripted computer agent produce different scores, but the universe of interest is how well a test taker can communicate with other persons, standardizing an assessment to include only interaction with a scripted computer agent may produce observed scores that do not generalize to the desired universe at all. In both of these cases, the observed score does not fully represent the construct of interest.

Estimating the variability in scores due to validity facets, such as through one or more generalizability studies, is a necessary step for making decisions about which conditions of a validity facet need to be included in an assessment to produce scores that generalize to the universe score of interest.

Other facets, which we term reliability facets, influence generalizability in a different way. Variability due to reliability facets influences the dependability of scores without necessarily changing (or restricting) the meaning of the construct measured. For example, suppose that a test taker's communication changes from one occasion to another (even when the group, the task, etc. are the same), but interest lies in generalizing over a wide range of occasions. Observing his communication on some occasions, but not others, may lead to questionable inferences to the universe score but should not affect the meaning of the observed score.

The solution is to include as many occasions as needed so that the average score across occasions generalizes to the universe score of interest. Including reliability facets in generalizability studies can show how many conditions of reliability facets are needed for dependable measurement.

Despite knowing the many sources of measurement error that may influence scores from assessments with groupwork, we do not know the magnitude of the error from potential sources. Consequently, we do not yet know, for example, how many standardized group compositions, or role structures, or task types, or occasions are needed for dependable measurement of groupwork skills and performance. Designing and carrying out generalizability studies will help inform these questions.

13.5 Additional Issues Introduced by the Use of Groupwork in Assessments

13.5.1 Relationships Between Process and Productivity/Performance

As described throughout this chapter, assessments with groupwork may produce scores for multiple constructs, some related to process, others related to productivity or performance. These scores may not be highly, or even positively, correlated. For example, while some studies find significant, and even high, correlations between group processes (e.g., providing mutual support, information exchange, communication, team initiative, and leadership) and quality of the team's performance or accuracy of decisions (Macmillan et al. 2013; Smith-Jentsch et al. 1998; Taggar and Brown 2001), others have reported weak or nonsignificant relationships between similar processes and outcomes (Meier et al. 2007; Chung et al. 1999). Still others produce conclusions in opposite directions, such as the density of adversarial relationships being positively or negatively related to team performance (Baldwin et al. 1997; Sparrowe et al. 2001).

One implication is that process and product/performance constructs are distinct and that measures of one cannot serve as proxies for the other. Another implication is that psychometric properties may vary for measures of process and product, and so may need to be investigated separately. For example, particular sources of measurement error may figure more prominently for some measures than for others (e.g., larger rater variation for measures of teamwork skills than for measures of group productivity). The optimal design for dependable measurement may, then, differ depending on the target construct.

13.5.2 Multiple Levels of Objects of Measurement: Individual and Group

Assessments with groupwork may yield scores at the individual-level (e.g., a test taker's teamwork skills, ability to communicate about the subject matter, performance during or after groupwork) and at the group level (e.g., the teamwork functioning of the group, the quality of the group's product or performance). The possibility of multiple levels of objects of measurement for assessments with groupwork introduces complexities not found in purely individual assessments. One is that reliability of scores may differ from one level to another. The relevant sources of measurement error, and hence the optimal design for reliable measurement, may not be the same or may not function in similar ways when, for example, measuring individuals' teamwork skills and when measuring the functioning of the group as a whole. For example, raters may find it more difficult to rate individual group members' contributions to resolving conflicts than to rate the group's success in conflict resolution, giving rise to lower rater agreement for individual scores than for group scores.

Conventional approaches for examining validity may also yield different results depending on whether the object of measurement is the individual or the group. That is, expert-novice comparisons (e.g., Brannick et al. 1995; Fowlkes et al. 1994; O'Neil et al. 1992; Smith-Jentsch et al. 1998), predictions of future performance (e.g., Arthur et al. 2003; Meriac et al. 2008; Speer et al. 2013), and examinations of the dimensionality of groupwork measures (for example, through exploratory or confirmatory factor analyses, O'Neil et al. 2010; Taggar and Brown 2001; or multitrait-multimethod analyses of the divergence of dimensions compared to the convergence of methods for measuring them, Brannick et al. 1993) may produce different results for individual and group scores. For example, giving suggestions and accepting suggestions may reflect teamwork skill dimensions that are more distinct (or separable) at the individual level than at the group level. These possibilities show that psychometric analyses need to attend to the particular unit of interest: individual, group, or both.

Another complexity is the statistical modeling issue arising from the non-independence of individuals within the group, especially when interest lies in producing dependable scores for individual examinees (on, say, an individual's ability to collaborate with others or engage in scientific argumentation). In collaborative settings, individuals' contributions are linked to, and dependent on, the contributions of other group members. Hence, the assumption of statistical independence of individuals' scores in conventional psychometric methods may not hold. New methods being explored for reliably scoring individuals' contributions to dynamic interactions during collaboration include dynamic factor analysis, multilevel modeling, dynamic linear models, differential equation models, social network analysis, intra-variability models, hidden Markov models, Bayesian belief networks, Bayesian knowledge tracing, machine learning methods, latent class analysis, neural networks, and point processes (von Davier and Halpin 2013; National Research Council 2013).

Another statistical modeling issue related to non-independence concerns the lack of independence from one group to another. Consider the desire to estimate the variability in individuals' scores across multiple group compositions. One way to gauge this variability is to observe the same individual in multiple groups that vary in terms of group composition attributes (e.g., ability level). A complexity arises when these groups have members in common beyond the target individual. As a consequence of the shared membership (which may be termed a multiple membership structure), groups are not independent. Multiple membership models (Goldstein 2003), which have been developed for similar situations such as estimating school effects when students attend more than one school (which might occur when a student changes schools mid-year), will be helpful here.

Yet another complexity introduced by the use of groupwork concerns task design. In an individual assessment, all work on a task comes from a single test taker. In assessments with groupwork, in contrast, one, some, or all members of a group may contribute to the task. If the intent is to measure collaborative skills (whether at the individual or group level), task designers must attend to features of the task that may inadvertently reduce opportunities for participation or communication with others. For example, easily divisible tasks or large and complex tasks may encourage groups to divide up the work and assign different group members different portions to complete, resulting in largely independent, rather than interactive, work. Similarly, tasks that can be completed by one person may also inhibit interaction among group members, albeit for different reasons. The desire to measure collaboration at the individual level poses an additional challenge: designing tasks that are likely to involve all, not just some, group members. In sum, then, task developers must be sensitive to possible consequences of task design for the nature and distribution of test takers' interaction on the task.

13.6 Conclusions

This chapter shows the immense complexity involved in arranging groupwork situations on assessments and the possible consequences for measuring groupwork processes, products, and performance. Multiple implications arise from this complexity. First, thought must be given to what outcome is being attributed to whom. The groupwork settings that are appropriate for assessment of individuals' participation in, or contributions to, groupwork may not be the same as those appropriate for assessment of the group's performance or productivity as a whole. Moreover, how the group functions may have different, and possibly conflicting, consequences for the measurement of individual participation and overall group performance. Second, it is important to consider multiple factors that may influence performance other than the skills of interest. For example, both individual participation and group performance may be influenced by the composition of the group, roles assigned to individual participants, the type of task, attributes of tasks of the same type, the occasion, and the rating process and rating scale. Obtaining reliable estimates of individual participation and group performance will require assessing individuals or groups across multiple instances of one or more of these factors. Third, designing a particular assessment will require close attention to these possible sources of score variation. The previous sections describe some strategies for addressing these factors, such as conducting generalizability studies to estimate the magnitude of sources of variation and using that information to make decisions about the design of assessments (e.g., the number of tasks or task types to be included, the number of different groupings to use for each test taker), or using avatars to represent group member attributes and behavior in an attempt to standardize groupwork experiences from one test taker to another or from one occasion to another.

Given the many sources of variation that potentially influence the measurement of processes and outcomes of groupwork, it is possible that the number of conditions needed for dependable measurement may simply be too large, especially when the object of measurement is the individual. Two alternative strategies for dealing with this issue are as follows:

Shift the focus from estimating the proficiency of the test taker (or small group) to estimating the performance at the level of classroom, school, or program. For example, even if the number of observations (e.g., tasks, groupings) is too small for dependable measurement of individuals (or of small groups), in some circumstances aggregating individual-level scores to higher levels may produce dependable measures of skills at the aggregate level. As described by Brennan (1995), we expect reliability of higher level units, such as schools, to be greater than reliability of lower level units, such as students, when the number of students sampled within schools is fairly large and variability of school means is large relative to the variability of students within schools.

Apply matrix sampling such that different groups within a classroom or school or program are assigned different conditions (e.g., different collections of tasks, different sets of group compositions). Matrix sampling may be an effective way to reduce the number of conditions per group and still maintain reliability at the aggregate level. Allowing different groups to be assigned to different conditions may also make it possible to handle a major challenge in assessments with groupwork–systematically manipulating group composition. Rather than making systematic assignments of particular group compositions to particular groups, which may not be feasible, an alternative is to form groups randomly. Doing so may help assure that a large number of group compositions will be represented across the classroom or school or program, and that effects associated with particular group compositions will cancel out in the aggregate.

Issues/Questions for Reflection

- Is the goal of the assessment to measure the contributions or performance of individual group members, the productivity or performance of the group as a whole, or both?
- What is the relative impact of potential sources of variation on the reliability of scores from the assessment with respect to the desired goal?
- How can the assessment be designed so that it incorporates the most important sources of score variation?
- How can the assessment be designed to produce reliable scores without sacrificing validity?

References

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–154.

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2010). Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire. *Quality and Safety in Health Care*, *19,* qshc–2009.036129.

Baldwin, T. T., Bedell, M. D., & Johnson, J. L. (1997). The social fabric of a team-based M.B.A. program: Network effects on student satisfaction and performance. *The Academy of Management Journal, 40*, 1369–1397.

Baron, J. B. (1992). SEA usage of alternative assessment: The Connecticut experience. In *Proceedings of the National Research Symposium on Limited English Proficient Student Issues* (2nd ed.). Washington, DC, September 4–6, 1991.

Baron, J. B. (1994). *Using multi-dimensionality to capture verisimilitude: Criterion-referenced performance-based assessments and the ooze factor*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April.

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *Journal of the Learning Sciences, 9*, 403–436.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1979/1980). Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2,* 191–218.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114–1124.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors, 37*, 641–651.

Brannick, M. T., Roach, R. M., & Salas, E. (1993). Understanding team performance: A multimethod study. *Human Performance, 6*, 287–308.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*, 385–396.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: Do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology, 95*, 506–523.

Chizhik, A. W., Alexander, M. G., Chizhik, E. W., & Goodman, J. A. (2003). The rise and fall of power and prestige order: Influence of task structure. *Social Psychology Quarterly, 66*, 303–317.

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15*, 463–493.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64*, 1–36.

Cohen, B. P., & Cohen, E. G. (1991). From groupwork among children to R&D teams: Interdependence, interaction, and productivity. *Advances in Group Processes, 8*, 205–225.

Collins, L. G., & Hartog, S. B. (2011). Assessment centers: A blended adult development strategy. In M. London (Ed.), *The Oxford handbook of lifelong learning* (pp. 231–250). New York, NY: Oxford University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *the dependability of behavioral measurements*. New York: Wiley.

Dorsey, D., Russell, S., Keil, C., Campbell, G., Van Buskirk, W., & Schuck, P. (2009). Measuring teams in action: Automated performance measurement and feedback in simulation-based training. In E. Salas, G. F. Goodwin & C. Shawn Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 351–381). New York, NY: Routledge.

Dyer, J. L. (2004), The measurement of individual and unit expertise. In J. W. Ness, V. Tepe & D. R. Ritzer (Eds.), *The science and simulation of human performance (advances in human performance and cognitive engineering research, volume 5)* (pp.11–124). Emerald Group Publishing Limited.

Esmonde, I. (2009). Mathematics learning in groups: Analyzing equity in two cooperative activity structures. *The Journal of the Learning Sciences, 18*, 247–284.

Fall, R., Webb, N. M., & Chudowsky, N. (2000). Group discussion and large-scale language arts assessment: Effects on students' comprehension. *American Educational Research Journal, 37*, 911–941.

Fay, D., Borrill, C., Amir, Z., Haward, R., & West, M. A. (2006). Getting the most out of multidisciplinary teams: A multi-sample study of team innovation in health care. *Journal of Occupational and Organizational Psychology, 79*, 553–567.

Ferrara, S. (2008). Design and psychometric considerations for assessments of speaking proficiency: The english language development assessment (ELDA) as illustration. *Educational Assessment, 13*, 132–169.

Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin & C. Shawn Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 411–431). New York, NY: Routledge.

Forsyth, C. M., Graesser, A. C., Pavlik, P., Cai, Z., Butler, H., Halpern, D. F., & Millis, K. (2013). Operation ARIES! Methods, mystery, and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining, 5*, 147–189.

Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs Methodology. *Military Psychology, 6*, 47–61.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. (1998). High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. *American Educational Research Journal, 35*, 225–267.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5,* 89–119.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Arnold.

Goodman, B. A., Linton, F. N., Gaimari, R. D., Hitzeman, J. M., Ross, H. J., & Zarrella, G. (2005). Using dialogue features to predict trouble during collaborative learning. *User Modeling and User-Adapted Interaction, 15*, 85–134.

Graesser, A., Rus, V., D'Mello, S., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. J. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 95–125). Charlotte, NC: Information Age Publishing.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer.

Guiton, G., Hodgson, C. S., Delandshere, G., & Wilkerson, L. (2004). Communication skills in standardized-patient assessment of final-year medical students: A psychometric study. *Advances in Health Sciences Education, 9*, 179–187.

Hoffman, B. J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise ratings. *International Journal of Selection and Assessment, 20*, 82–97.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology, 1*, 98–104.

Johnson, W. L. (2010). Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education, 20*, 175–195.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681–706.

Kenderski, C. M. (1983). *Interaction processes and learning among third grade Black and Mexican-American students in cooperative small groups*. Ph.D. Dissertation, University of California, Los Angeles.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free rider effects. *Journal of Personality and Social Psychology, 44*, 78–94.

Kilduff, M., Crossland, C., Tsai, W., & Krackhardt, D. (2008). Organizational network perceptions versus reality: A small world after all? *Organizational Behavior and Human Decision Processes, 107,* 15–28.

Kumbasar, E., Kimball Romney, A., & Batchelder, W. H. (1994). Systematic biases in social perception. *American Journal of Sociology, 100*, 477–505.

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13*, 435–455.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*, 151–172.

Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology, 41*, 585–634.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247–258.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1–16.

Loughry, M. L., Ohland, M. W., & Moore, D. D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement, 67*, 505–524.

MacMillan, J., Entin, E. B., Morley, R., & Bennett, W. (2013). Measuring team performance in complex and dynamic military environments: The SPOTLITE Method. *Military Psychology, 25*, 266–279.

Malone, M. E. (2003). Research on the oral proficiency interview: Analysis, synthesis, and future directions. *Foreign Language Annals, 36*, 491–497.

Maryland State Department of Education. (1994). *Maryland school performance assessment program: Public release tasks*. Baltimore, MD: Maryland State Department of Education.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*, 273–283.

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management, 40*, 130–160.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–421.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *Computer-Supported Collaborative Learning, 2*, 63–86.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052.

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS system: Reliability and validity of a tool for measuring teamwork behavior in the operating theatre. *Quality and Safety in Health Care, 18*, 104–108.

Morgan, P. J., Pittini, R., Regehr, G., Marrs, C., & Haley, M. F. (2007). Evaluating teamwork in a simulated obstetric environment. *Anesthesiology*, *106*, 907–915.

Mu, J., Stegmann, K., Mayfield, E., Rose, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *Computer-Supported Collaborative Learning, 7*, 285–305.

Mulryan, C. (1992). Student passivity during cooperative small groups in mathematics. *Journal of Educational Research, 85*, 261–273.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*, 483–508.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core standards mathematics*. Washington, DC: Authors.

National Research Council. (2011). *Assessing 21st century skills*. Washington, DC: National Academies Press.

National Research Council. (2013). *New directions in assessing performance of individuals and groups: Workshop summary*. Washington, DC: National Academies Press.

Neuberger, W. (1993, September). *Making group assessments fair measures of students' abilities*. Paper presented at the National Center for Research on Evaluation, Standards, and Student Testing's Conference on "Assessment Questions: Equity Answers", UCLA, Los Angeles, CA.

O'Donnell, A. M. (1999). Structuring dyadic interaction through scripted cooperation. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 179–196). Hillsdale, NJ: Erlbaum.

O'Neil, H. F., Allred, K., & Dennis, R. (1992). *Simulation as a performance assessment technique for the interpersonal skill of negotiation*. Technical report, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

O'Neil, H. F., Chuang, S. S., & Baker, E. L. (2010). Computer-based feedback for computer-based collaborative problem solving. In D. Ifenthaler et al. (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (pp. 261–279). Springer Science + Business Media.

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*, 292–308.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., et al. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning and Education, 11*, 609–630.

Organisation for Economic Co-operation and Development (2013, March). *PISA 2015 draft collaborative problem solving framework.* Available: www.oecd.org

Pomplun, M. (1996). Cooperative groups: Alternative assessment for students with disabilities? *The Journal of Special Education, 30*, 1–17.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan,v?,* 146–155.

Richter Lagha, R. (2013). *Accuracy of professional self-reports.* Ph.D. Dissertation, University of California, Los Angeles.

Richter Lagha, R., Boscardin, C. K., May, W., & Fung, C. C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine, 87*, 1077–1082.

Rimor, R., Rosen, Y., & Naser, K. (2010). Complexity of social interactions in collaborative learning: The case of online database environment. *Interdisciplinary Journal of E-Learning and Learning objects, 6*, 355–365.

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuel, B., Nussbaum, M., & Claro, S. (2009). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. Education Technology Research Development. doi:10.1007/211423-009-9142-9

Rose, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Computer-Supported Collaborative Learning, 3*, 237–271.

Rosen, Y., & Tager, M. (March, 2013). *Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach*. Research report, Pearson.

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: Feasibility of clinical and nonclinical assessor calibration with short-term training. *Annals of Surgery, 255*, 804–809.

Salas, E., Burke, C. S., & Cannon-Bowers, J. A. (2000). Teamwork: Emerging principles. *International Journal of Management Reviews, 2*, 339–356.

Salas, E., Burke, C. S., & Cannon-Bowers, J. A. (2003). Teamwork: Emerging principles. *International Journal of Management Reviews, 2*, 339–356.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "Big Five" in teamwork? *Small Group Research, 36*, 555–599.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*, 89–99.

Sampson, V., & Clark, D. B. (2011). A comparison of the collaborative scientific argumentation practices of two high and two low performing groups. *Research in Science Education, 41*, 63–97.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994) *The effects of working in pairs in science performance assessments*. Santa Monica, CA: The Rand Corporation. Manuscript submitted for publication.

SCANS (1999). *Skills and Tasks for Jobs: A SCANS Report for 2000.* Washington, DC: U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills (SCANS).

Schellens, T., Van Keer, H., & Valcke, M. (2005). The impact of role assignment on knowledge construction in asynchronous discussion groups: A multilevel analysis. *Small Group Research, 36*, 704–745.

Schoor, C., & Bannert, M. (2011). Motivation in a computer-supported collaborative learning scenario and its impact on learning activities and knowledge acquisition. *Learning and Instruction, 21*, 560–573.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42–54.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood Cliffs, NJ: Prentice-Hall.

Smith-Jentsch, K. A., Cannon-Bowers, J. A., Tannenbau, S. I., & Salas, E. (2008). Guided team self-correction: Impacts on team mental models, processes, and effectiveness. *Small Group Research, 39*, 303–327.

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998). Measuring team-related expertise in complex environments. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (Vol. 1, pp. 61–87). Washington, DC: American Psychological Association.

Smith-Jentsch, K. A., Salas, E., & Baker, D. P. (1996). Training team performance-related assertiveness. *Personnel Psychology, 49*, 909–936.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. Santa Monica, CA: RAND Corporation.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *The Academy of Management Journal, 44*, 316–325.

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2013). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology.* doi:10.1037/a0035213

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science, 40*, 297–323.

Steiner, I. (1972). *Group process and productivity*. New York, NY: Academic Press.

Taggar, S., & Brown, T. C. (2001). Problem-solving team behaviors: Development and validation of BOS and a hierarchical factor structure. *Small Group Research, 32*, 698–726.

Von Davier, A. A., & Halpin, P. F. (2013, December). Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations. Research Report ETS RR-13-41. Educational Testing Service.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds., Trans.). Cambridge, MA: Harvard University Press.

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology, 24*, 108–124.

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks,*. doi:10.1016/j.socnet.2012.01.003

Webb, N. M. (1984). Stability of small group interaction and achievement over time. *Journal of Educational Psychology, 76*, 211–224.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1*, 131–152.

Webb, N. M., & Mastergeorge, A. M. (2003). The development of students' learning in peer-directed small groups. *Cognition and Instruction, 21*, 361–428.

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science performance. *American Educational Research Journal, 39*, 943–989.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Macmillan.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior, 26*, 506–515.

Wildman, J. L., Thayer, A. L., Rosen, M. A., Salas, E., Mathieu, J. E., & Rayne, S. R. (2012). Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Review, 11*, 97–129.

Wimmers, P., & Lee, M. (2014). Identifying longitudinal growth trajectories of learning domains in problem-based learning: A latent growth curve modeling approach using SEM. *Advances in Health Sciences Education.* doi:10.1007/s10459-014-9541-5

Wise, N., & Behuniak, P. (1993, April). *Collaboration in student assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Zhang, B., & Ohland, M. W. (2009). How to assign individualized scores on a group project: An empirical evaluation. *Applied Measurement in Education, 22*, 290–308.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the Learning Sciences, 18*, 7–44.

Chapter 14
Evaluating Team Performance: A Systematic Review

Danette W. McKinley

Abstract Effective teamwork amongst healthcare professionals has been shown to correlate with positive patient outcomes. This paper reviews research conducted with healthcare professionals to determine the extent to which assessments of team performance had been developed and evaluated between 2006 and 2012. The National Library of Medicine's indexed database PubMed was used to identify potential articles for inclusion in the review. Of the 549 articles retrieved, 158 articles were selected for inclusion in the study based on review of the article abstracts. Of the 158 articles, 26 of the articles examined psychometric characteristics of the measures. Most instruments were observation checklists, and research was conducted primarily in emergency medicine and surgery. Measures developed that can be used in a variety of healthcare settings, in addition to surgery and acute care, will be invaluable as the complexity of providing adequate patient care will increasingly require the coordinated efforts of team members.

Takeaways

- Frameworks for training health professionals in functioning as teams should provide the basis for the development of assessment instrucments.
- In designing an assessment of skills in teamwork, one should consider how the results are going to be used. Is the intention to provide feedback to individual team members? Will assessment results be used to evaluate educational programmes?.
- Assessment purpose will determine content and format for the assessment.
- A number of instruments have already been developed and validated. If at all possible, use an assessment that has been studied.

D.W. McKinley (✉)
Foundation for Advancement of International Medical Education and Research,
Philadelphia, PA, USA
e-mail: dmckinley@faimer.org

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change
in Professional Education 13, DOI 10.1007/978-3-319-30064-1_14

14.1 Evaluating Team Performance: A Systematic Review

Almost all of us are familiar with teams and/or team observation thanks to sports teams. Whether recreational, school-based, or professional, athletic teams are an example of what a group of people can do to reach a common goal. Assessment of team performance in sports is well established, and the outcomes are obvious (did they win?). For teams in aviation and health care, the outcomes are significantly more important. Did the plane land safely? Is the patient alive? Aviation and health care share a common feature, the potential of tragic consequences when errors occur. Because of the consequences associated with failure, these professions are said to be examples of high-reliability organizations.

The nature of the "hypercomplex environment" in which health care occurs is characterized by several decision makers, whose roles are embedded in an "extreme hierarchical differentiation," they note that the assurance of patient safety requires interaction and communication in "compressed time" with a "high degree of accountability" (Baker et al. 2006). In identifying the characteristics of high-reliability organizations, Baker, Day, and Salas argue that healthcare providers are often organized in teams, and that their interactions are part of the vital operations in various settings (Baker et al. 2006). The hypercomplexity of the context in which health care occurs is characterized by specialization, where team members have specific roles, responsibilities, and knowledge (Orchard et al. 2012). Because errors, although rare, result in serious consequences, teamwork is essential. Knowledge of their own roles and responsibilities, monitoring of team member performance, and a positive attitude toward teamwork have been shown to relate to team effectiveness (Baker et al. 2006; Driskell and Salas 1992).

Team competencies typically considered for training programs have been identified as leadership, mutual performance monitoring, mutual support, adaptability, team orientation, mutual trust, shared mental models, and communication (Baker et al. 2006; Salas et al. 2005). The extent to which interdependent healthcare professionals are able to communicate and coordinate activities has been shown to relate to failures as well as successes (Healey et al. 2006; Nagpal et al. 2012). While team competencies have been studied extensively in aviation and the military, research on teamwork training and evaluation is gaining prominence in healthcare professions (e.g., Capella et al. 2010; Lurie et al. 2011; Tumerman and Carlson 2012). When considering patient outcomes, patient safety and team performance have been linked to surgery (e.g., Lawton et al. 2012; Nagpal et al. 2012).

Considerable research has been conducted in organizational and cognitive psychology, providing a theoretical basis for competencies associated with effective team performance. While several training programs focusing on team training have been started in health professions, little research has been done to determine whether the theories developed for aviation and military operations are applicable to healthcare settings (Baker et al. 2006). The increasing emphasis on the link between effective teamwork and positive patient outcomes demands that research provide evidence supporting training programs as well as assessment and evaluation of team processes in health care.

Measurement of team processes that lead to successful performance can be challenging. In the health professions, teamwork training introduces new concepts, where autonomy had previously been emphasized (Lerner et al. 2009). Research in other fields has shown that the development of readily observable behavior checklists is likely to be more accurate than self-assessment (Baker and Salas 1992), although more work regarding the extent to which these measures are specialty- or context-specific needs to be done. What is the current state of affairs in assessment of teamwork for these professionals? How has theory from other fields been incorporated in the training and evaluation of health professionals?

Principles meant to guide the development of assessment tools provide a framework for the categorization of the research reviewed in this paper. Baker and Salas (1992) indicated that measures should show clear relation to theory, should present evidence of reliability and validity, should indicate the developmental nature of teams, and should be observation based. In the current investigation, a review of research published that involved healthcare professionals was conducted to determine the extent to which assessments of teamwork had been developed and evaluated between 2006 and 2012. The goal of the study was to summarize the extent to which instruments had been developed, adhering to the principles proposed by Baker and Salas (1992).

14.2 Methods

The National Library of Medicine's indexed database PubMed was used to identify potential articles for inclusion in the review. Articles published in English between 2006 and 2012 were searched for potential inclusion. Using the search term "assessment OR evaluation AND teamwork," a total of 549 articles were retrieved. Only those articles that referenced assessment of teamwork were selected for inclusion. Articles were classified by healthcare profession (e.g., pharmacy, medicine, nursing), specialty (e.g., surgery, oncology, anesthesiology), type of research (e.g., program evaluation, quality assurance, psychometric), and factor measured (process, skill, or task).

Of the 549 articles retrieved, 158 articles were selected for inclusion in the study based on review of the article abstracts. Since the current investigation focuses on assessment and evaluation of teamwork amongst health professionals, articles focusing on training, quality assurance, or safety climate were eliminated from further review. Of the 158 articles included for second-level review, 45 were identified that focused specifically on assessment or evaluation of teams, the remainder focused on some other aspect of teamwork (see Fig 14.1).

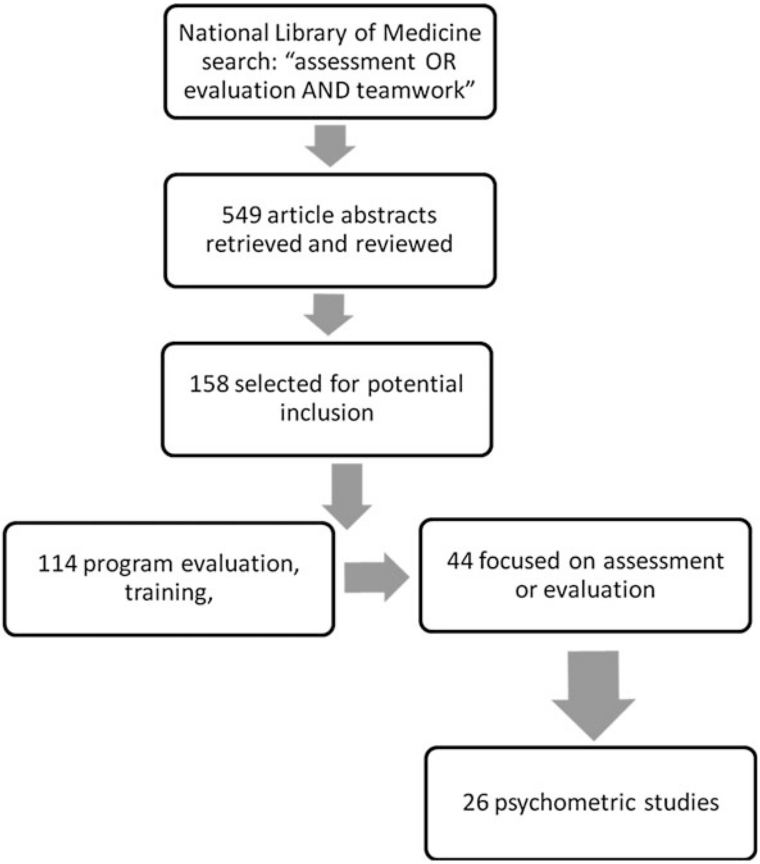

National Library of Medicine search: "assessment OR evaluation AND teamwork"
549 article abstracts retrieved and reviewed
158 selected for potential inclusion
114 program evaluation, training,
44 focused on assessment or evaluation
26 psychometric studies

Fig. 14.1 Review and selection process

14.2.1 Results

Nonmeasurement articles on teamwork. Based on the review of 158 articles on teamwork, 113 of the articles reported research focused on topics other than assessment, most of which reported on program evaluation (45 articles, 40 %). Although the majority of these studies focused on medicine (n = 53, 47 %), a number of the studies included various members of the healthcare team (48 articles, 42 %). While surgical teams (n = 12; 11 %) and emergency medicine teams (n = 8; 7 %) were studied, there was an effort to study healthcare teams in a variety of specialties (n = 48; 42 % of articles did so). The primary focus of the articles was on program or training evaluation (n = 45; 40 %), theory specific to healthcare settings (n = 15; 13 %), and program development (n = 15; 13 %). The remaining categories included quality assurance and patient safety (10 articles), review of research on teamwork (9 articles), and attitudes toward teamwork (5 articles), amongst others. Appendix A provides the listing of articles on teamwork that were not focused on assessment.

Review of the 45 articles on assessment and evaluation was primarily focused on psychometric issues (i.e., reliability and/or evidence of validity); 26 of the articles examined psychometric characteristics of the measures (59 %). Other study types included evaluation (*n* = 11; 22 %), theoretical study of teamwork competencies specific to healthcare professionals (4 articles, 9 %), and review articles on teamwork measures in health professions (*n* = 2; 4 %). Most of the studies were focused on medicine (*n* = 27; 60 %), predominately in surgery (*n* = 17; 37 %), although several of the studies that reported on measures were interdisciplinary (i.e., across professions, but in a particular specialty). Articles classified as general (*n* = 12; 28 %) included settings that crossed specialties (e.g., Orchard et al. 2012). The investigations primarily focused on measures intended to evaluate teamwork skills (*n* = 22; 49 %), with eight (18 %) of the measurement themed articles measuring skills and tasks, and eight measuring only tasks. The remaining articles measured attitudes toward teamwork (*n* = 5; 11 %), climate (*n* = 1; 2 %), and the relation between teamwork and patient outcomes (1 article).

Because the focus of the article is specifically on assessment, further characterization of articles reporting on measures that examined psychometric qualities were conducted to determine whether the measures studied were self-report, self-assessment, or meant for observation. Of the 26 articles containing information specific to the measure(s) studied, one did not provide sufficient information to determine how the measure would be used (Varkey et al. 2009). Of the remaining 25 articles, 9 (35 %) were self-report or self-assessment measures; and two of these were attitude toward teamwork measures. The remaining 17 articles were observational measures; 8 of those were conducted in surgical settings. Generally, team activities were videotaped, and then the checklists were used to rate performances. Table 14.1 provides information on the articles in which measures were studied.

When examining which competencies were measured, the work of Salas et al. 2008 was used as the theoretical basis for team training. These included team leadership, mutual performance monitoring, backup behavior, adaptability, team orientation, shared mental models, mutual trust, and closed-loop communication (Salas et al. 2008, p. 1003). Whether these constructs were measured as part of the assessment was examined by review of articles that detailed instrument content. Table 14.2 presents the overlap between the theorized constructs and those measured in the studies included in the review. The work of Patterson et al. (2012) showed that it was possible to design an instrument that measured all of the competencies for effective teamwork, and in addition, they included measures of conflict. Lurie et al. 2011 studied whether the burden of rating using longer checklists could be reduced without loss of information and reliability. In their study, they found that a 29-item checklist could be reduced to as few as five items with similar reliability and factor structure, and that observations could be completed in as few as 3 minutes or less.

For articles that included the measures as an appendix, authors often found that the items used could be labeled as constructs other than those included in theory focused on training. Team orientation, shared mental models, and mutual trust may have been measured in studies of attitudes more often than in studies which focused on evaluation of teamwork skills. For the six studies included in Table 14.2, most included measures of communication and leadership. These factors are those

Table 14.1 Studies focused on assessment of teamwork

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2006	David R. King	Simulation training for a mass casualty incident: two-year experience at the Army Trauma Training Center	Evaluation	Medicine	Emergency/Trauma	Meant to identify skills for additional training
2006	E.J. Thomas	Teamwork and quality during neonatal care in the delivery room	Evaluation	Interdisciplinary	Neonatology	Measure used with videotaped teamwork in neonatal resuscitation. Intended to use results to improve training
2006	JoDee M. Anderson	Simulating extracorporeal membrane oxygenation emergencies to improve human performance. Part II: assessment of technical and behavioral skills	Evaluation	Medicine	Emergency/Trauma	Nursing simulation based on very specific type of emergency. Pretest–posttest design to evaluate training program. Observation-based measure
2006	Katherine C. Pollard	A comparison of interprofessional perceptions and working relationships among health and social care students: the results of a 3-year intervention	Evaluation	Interdisciplinary	General	Participant perception of how team members related to each other
2006	Susan Mann	Assessing quality obstetrical care: development of standardized measures	Evaluation	Medicine	Obstetrics	Teamwork measured as part of a study on quality indicators for obstetrics. Benchmarking study
2006	A. Hutchinson	Use of a safety climate questionnaire in UK health care: factor structure, reliability and usability	Psychometric	Interdisciplinary	General	Safety climate survey
2006	Aysegul Yildirim	Turkish version of the Jefferson Scale of Attitudes Toward Physician-Nurse Collaboration: a preliminary study	Psychometric	Nursing	General	Nurses and physicians rate attitudes about collaboration between the two groups

(continued)

Table 14.1 (continued)

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2006	J. Bryan Sexton	Teamwork in the operating room: frontline perspectives among hospitals and operating room personnel	Psychometric	Medicine	Surgery	
2006	Steven Yule	Development of a rating system for surgeons' non-technical skills	Psychometric	Medicine	Surgery	
2006	Shabnam Undre	Observational assessment of surgical teamwork: a feasibility study	Psychometric	Medicine	Surgery	Used OTAS (see Healey)
2006	Andrew N. Healey	The complexity of measuring interprofessional teamwork in the operating theatre	Theory	Medicine	Surgery	
2007	Shabnam Undre	Multidisciplinary crisis simulations: the way forward for training surgical teams	Evaluation	Interdisciplinary	Surgery	
2007	Daniel L. Davenport	Risk-adjusted morbidity in teaching hospitals correlates with reported levels of communication and collaboration on surgical teams but not with scale measures of teamwork climate, safety climate, or working conditions	Organizational climate	Medicine	Surgery	
2007	Allan Frankel	Using the communication and teamwork skills (CATS) Assessment to measure health care team performance	Psychometric	Interdisciplinary	General	
2008	Mirjam Körner	Analysis and development of multiprofessional teams in medical rehabilitation	Psychometric	Medicine	Physical medicine and rehabilitation	

(continued)

Table 14.1 (continued)

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2008	Steven Yule	Surgeons’ non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system	Psychometric	Medicine	Surgery	
2008	Amy H. Kaji	Assessing hospital disaster preparedness: a comparison of an on-site survey, directly observed drill performance, and video analysis of teamwork	Quality assurance	Interdisciplinary	General	Disaster preparedness
2008	K. Catchpole	Teamwork and error in the operating room: analysis of skills and roles	Theory	Interdisciplinary	Surgery	
2009	P. McCulloch	The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre	Evaluation	Interdisciplinary	Surgery	Crew resource management training applied to healthcare setting. Pretest–posttest evaluation of training with observation of skills. Used to identify areas challenges based on measures for specific types of surgical procedures
2009	A. Mishra	The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre	Psychometric	Interdisciplinary	Surgery	
2009	Karen Mazzocco	Surgical team behaviors and patient outcomes	Psychometric	Interdisciplinary	Surgery	
2009	Melanie C. Wright	Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance	Psychometric	Medicine	General	

(continued)

Table 14.1 (continued)

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2009	Nicholas Hamilton	Team behavior during trauma resuscitation: a simulation-based performance assessment	Psychometric	Medicine	Emergency/Trauma	
2009	Nick Sevdalis	Observational teamwork assessment for surgery: construct validation with expert versus novice raters	Psychometric	Medicine	Surgery	
2009	Prathibha Varkey	An innovative team collaboration assessment tool for a quality improvement curriculum	Psychometric	Interdisciplinary	General	
2009	D. Tregunno	Development and usability of a behavioural marking system for performance assessment of obstetrical teams	Theory	Interdisciplinary	Obstetrics	
2010	David P. Baker	Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire	Attitudes	Interdisciplinary	General	Relationship between attitude toward teamwork and team functioning
2010	Jeannette Capella	Teamwork training improves the clinical care of trauma patients	Evaluation	Medicine	Emergency/Trauma	Reported on the evaluation of team training using “Trauma team performance observation tool”
2010	Chris Kenaszchuk	Validity and reliability of a multiple-group measurement scale for interprofessional collaboration	Psychometric	Interdisciplinary	General	Nurses rating physician skills
2010	Simon Cooper	Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM)	Psychometric	Medicine	Emergency/Trauma	

(continued)

Table 14.1 (continued)

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2010	Michael A. Rosen	Tools for evaluating team performance in simulation-based training	Review	Medicine	Emergency/Trauma	Review article on instruments evaluating team performance
2011	Jan Schraagen	Assessing and improving teamwork in cardiac surgery	Evaluation	Medicine	Surgery	Coding of nonroutine events in the OR: "Real-time teamwork observations were supplemented with process mapping, questionnaires on safety culture, level of preparedness of the team, difficulty of the operation, and outcome measures"
2011	Jennifer Weller	Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams	Psychometric	Medicine	Critical care	Measure of individuals within team
2011	Louise Hull	Assessment of stress and teamwork in the operating room: an exploratory study	Psychometric	Medicine	Surgery	Relationship between stress and teamwork in the OR
2011	Stephen J. Lurie	Assessing teamwork: a reliable five-question survey	Psychometric	Medicine	Family medicine	
2011	Bharat Sharma	Non-technical skills assessment in surgery	Review	Medicine	Surgery	Review article on instruments evaluating team performance
2012	Jeffrey Braithwaite	A four-year, systems-wide intervention promoting interprofessional collaboration	Evaluation	Interdisciplinary	General	Measure of attitudes toward interprofessional collaboration and learning
2012	Simon Cooper	Managing patient deterioration: assessing teamwork and individual performance	Evaluation	Nursing	Emergency/Trauma	Reported using the "Team emergency assessment measure" and observation during OSCE to assess teamwork. Program evaluation more than assessment of teams

(continued)

Table 14.1 (continued)

Publication Year	First Author	Title	Study type	Profession	Specialty	Comments
2012	C. Taylor	Developing and testing TEAM (Team Evaluation and Assessment Measure), a self-assessment tool to improve cancer multidisciplinary teamwork	Psychometric	Medicine	Oncology	Self-assessment measure
2012	Carole A. Orchard	Assessment of interprofessional team collaboration scale (AITCS): development and testing of the instrument	Psychometric	Interdisciplinary	General	Individual measure of perception of collaboration amongst team members including patients and their families
2012	Douglas R. Wholey	The teamwork in assertive community treatment (TACT) scale: development and validation	Psychometric	Medicine	Psychiatry	
2012	Kamal Nagpal	Failures in communication and information transfer across the surgical care pathway: interview study	Psychometric	Medicine	Surgery	Study of reliability and feasibility of a tool measuring teamwork based on postoperative handover
2012	Kevin J. O'Leary	Assessment of teamwork during structured interdisciplinary rounds on medical units	Psychometric	Medicine	General	Adaptation of OTAS instrument demonstrated relevance across settings
2012	P. Daniel Patterson	Measuring teamwork and conflict among emergency medical technician personnel	Psychometric	Medicine	Emergency/Trauma	Measure of task performance for EMTs
2012	Stephanie Russ	Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training	Psychometric	Medicine	Surgery	

Table 14.2 Constructs measured by teamwork assessments

Publication year	First author	Title	Profession	Leadership	Communication	Mutual performance monitoring	Backup behavior	Adaptability	Team orientation	Shared mental models	Mutual trust	Other
2010	Jeannette Capella	Teamwork training improves the clinical care of trauma patients	Medicine	Y	Y	Y	Y	N	N	N	N	
2011	Louise Hull	Assessment of stress and teamwork in the operating room: an exploratory study	Medicine	Y	Y	Y	Y	N	N	N	N	Coordination
2010	Chris Kenaszchuk	Validity and reliability of a multiple-group measurement scale for interprofessional collaboration	Interdisciplinary	N	Y	N	Y	N	N	N	N	Isolation
2011	Stephen J. Lurie	Assessing teamwork: a reliable five-question survey	Medicine	Y	Y	N	Y	N	Y	Y	Y	Problem-solving
2012	P. Daniel Patterson	Measuring Teamwork and Conflict among Emergency Medical Technician Personnel	Medicine	Y	Y	Y	Y	Y	Y	Y	Y	Process conflict; task conflict; interpersonal conflict
2012	Carole A. Orchard	Assessment of Interprofessional Team Collaboration Scale (AITCS): development and testing of the instrument	Interdisciplinary	Y	Y	N	N	N	Y	N	N	Partnership

considered to have the most negative effects when teamwork fails (Nagpal et al. 2012). Two studies (Kenaszchuk et al. 2010; Patterson et al. 2012) considered negative behaviors that could hamper teamwork, isolation, and sources of conflict. In general, instruments described were theory based, and the authors defined constructs in a manner consistent with the theoretical literature.

14.3 Discussion

Research conducted with the military and aviation has informed practices in health care (Baker et al. 2006; Kendall and Salas 2004; Salas et al. 2008). Research on the development of measures that are reliable and based on theory have been conducted and have advanced adaptation of measures of team interaction from other professions. The predominance of articles on psychometric issues is not surprising; work to determine whether measures could be developed or adapted for use with healthcare teams in a variety of settings is essential, and was recommended by those pivotal in the development of teamwork theory in other professions (e.g., Baker and Salas 1992). In addition, study of the factors that can affect teamwork or result in negative performance (e.g., Baker 2010; Capella et al. 2010) support the identification of factors related to the avoidance of negative consequences.

Measurement of teamwork amongst healthcare professionals faces several challenges. First, efforts continue to be specialty-specific (e.g., surgical, emergency medicine, community medicine), although there are studies that have looked to see if the measures can be used across settings (e.g., O'Leary et al. 2012). While several of the measures developed are based in theory, different constructs may be measured. Although there was minimal inconsistency in terminology, papers that do not clearly define the constructs measured can make this effort challenging, particularly if measures are to be used across health professionals and settings.

Interestingly, research has shown that team members are generally not reliable at assessing their level of skill (Baker and Salas 1992; Eva et al. 2004), but practitioners are generally able to self-monitor (Eva and Regehr 2010). Seven of the articles reviewed were self-report or self-assessment measures. Since two of those were attitude measures, the importance of the effect of self-assessment may not have the same significance as it does in measurement of competence.

Although observational measures have been said to be preferable, securing the necessary number of raters to produce reliable measures has been challenging (Morgan et al. 2007), although recent work has shown promise (Russ et al. 2012). Efforts are underway to show that shorter versions of long measures can be used in a fashion that may facilitate recruitment and training of raters, generating more ratings available for the evaluation of teamwork skills (Lurie et al. 2011).

A number of publications that focused on program evaluation highlight a challenge in assessment of teams: finding reliable measurement tools that assess group interaction (Morgan et al. 2007; Murray and Enarson 2007), particularly when targeted training in teamwork skills has been conducted. These studies typically rely on pretest–posttest design (e.g., Aboumatar et al. 2012; Vyas et al. 2012), and often include measures of participant opinion regarding training. While this is legitimate for program evaluation, measures that can be used in practice (i.e., workplace setting) can provide additional evidence of the effect of training in interprofessional teamwork.

Although this review has provided information on the development of assessment tools for use with healthcare professionals it is not without limitations. First, only the author reviewed the abstracts, so no consistency of coding was provided. Other researchers may disagree with the categorization of the studies included in this report. However, the appendices, which include a complete reference list, can be used by others who are interested in the topic. Also, only abstracts were reviewed to determine inclusion/exclusion, and reference lists of included articles were not used to identify other articles for potential inclusion. Additional review and categorization may have examined whether reliability was reported, and the extent to which validity evidence was provided for the measures. Despite these limitations, this review provides preliminary information on the methods used to evaluate teamwork amongst healthcare professionals.

The nature of health care, as typified by Baker et al. 2006, is increasingly complex, and errors have serious consequences. The rigid hierarchical roles that health professionals have traditionally had must change; although knowledge of each person's role in the team is essential, adaptability and monitoring are important components of successful teamwork. Studies have begun to show the relationship between effective teams and positive patient outcomes (e.g., Mazzocco et al. 2009). Measures developed that can be used in a variety of healthcare settings, in addition to surgery and acute care, will be invaluable as the complexity of providing adequate patient care will increasingly require the coordinated efforts of team members.

Issues/Questions for Reflection

- Training in use of the assessment may be necessary, particularly if observation of teams will occur.
- How can work done in psychology on human interaction support the assessments developed for teamwork?
- The effect of team size, team formation (standing teams vs. dynamic teams), hierarchical structure, professional identity, and more will need to be studied in compiling evidence of validity of team measures.

Appendices

Appendix 1: Studies Without Investigation of Teamwork Measure

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2006	A. Flabouris	Incidents during out-of-hospital patient transportation	Quality assurance	Medicine	General	Evaluation of factors affecting adverse outcomes in patient transport
2006	Alison Bellamy	Case reviews: promoting shared learning and collaborative practice	Evaluation	Medicine	General	Case review of teamwork
2006	B.J. Moran	Decision-making and technical factors account for the learning curve in complex surgery	Clinical	Medicine	Surgery	
2006	Cheryl Knapp	Bronson Methodist Hospital: journey to excellence in quality and safety	Evaluation	Medicine	General	QA for Hospital
2006	Debra Parker Oliver	Inside the interdisciplinary team experiences of hospice social workers	Evaluation	Medicine	Social work	Study of experiences of social workers in hospice care
2006	D. Lamb	Collaboration in practice—assessment of an RAF CCAST	Theory	Medicine	Critical care	Looked at the factors in critical care that may affect climate
2006	E. Anderson	Evaluation of a model for maximizing interprofessional education in an acute hospital	Program development	Interdisciplinary	Emergency/Trauma	Development and evaluation of a training program in an acute care setting
2006	E.K. Mayer	Robotic prostatectomy: the first UK experience	Procedure evaluation	Medicine	Urology	Clinical procedure evaluation, not team assessment
2006	Eileen B. Entin	Training teams for the perioperative environment: a research agenda	Theory	Medicine	Surgery	
2006	Elaine Cole	The culture of a trauma team in relation to human factors	Theory	Interdisciplinary	Emergency/Trauma	Ethnographic study of trauma team culture

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2006	Elie A. Akl	Brief report: Internal medicine residents', attendings', and nurses' perceptions of the night float system	Evaluation	Medicine	Internal medicine	Evaluation of residents on night shift
2006	H. Patrick McNeil	An innovative outcomes-based medical education program built on adult learning principles	Training	Medicine	General	
2006	J. Randall Curtis	Intensive care unit quality improvement: a "how-to" guide for the interdisciplinary team	Quality assurance	Interdisciplinary	Critical care	
2006	James K. Takayesu	How do clinical clerkship students experience simulator-based teaching? A qualitative analysis	Evaluation	Medicine	General	
2006	Jean Kipp	What motivates managers to coordinate the learning experience of interprofessional student teams in service delivery settings?	Evaluation	Interdisciplinary	Various	
2006	Jill Scott-Cawiezell	Nursing home safety: a review of the literature	Review	Nursing	Geriatrics	Author argued that "better outcome measures must be developed that are nurse sensitive"
2006	Kanakarajan Saravanakumar	The challenges of obesity and obstetric anaesthesia	Clinical	Medicine	Anesthesiology	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2006	Karen S. Martin	Introducing standardized terminologies to nurses: Magic wands and other strategies	Program Development	Nursing	General	Focused on nurses' of clinical data information collection
2006	Kathleen Rice Simpson	Nurse-physician communication during labor and birth: implications for patient safety	Theory	Interdisciplinary	Obstetrics	Description of communication between nurses and physicians with suggestions for improvement to teamwork for improved patient safety
2006	Kenn Finstuen	Executive competencies in healthcare administration: preceptors of the Army-Baylor University Graduate Program	Theory	Medicine	General	Preceptor competencies
2006	Marsha Sharp	Enhancing interdisciplinary collaboration in primary health care	Report	Interdisciplinary	Dieticians	
2006	Martin Rhodes	Teaching evidence-based medicine to undergraduate medical students: a course integrating ethics, audit, management and clinical epidemiology	Evaluation	Medicine	General	
2006	Nadia Abdulhadi	Quality of interaction between primary health-care providers and patients with type 2 diabetes in Muscat, Oman: an observational study	Evaluation	Medicine	Primary care	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2006	Phillip G. Clark	What would a theory of interprofessional education look like? Some suggestions for developing a theoretical framework for teamwork training 1	Theory	Interdisciplinary	General	
2006	Ping-Chuan Hsiung	Evaluation of inpatient clinical training in AIDS care	Program development	Medicine	General	Medical students’ attitudes about AIDS
2006	Roy T. Dobson	Interprofessional health care teams: attitudes and environmental factors associated with participation by community pharmacists	Theory	Pharmacy	General	Participation of pharmacists as members of primary healthcare team
2006	S.M. Handler	Patient safety culture assessment in the nursing home	Patient safety	Interdisciplinary	Geriatrics	
2006	S. Yule	Non-technical skills for surgeons in the operating room: a review of the literature	Theory	Medicine	Surgery	
2006	Sally O. Gerard	Implementing an intensive glucose management initiative: strategies for success	Evaluation	Nursing	Primary care	
2006	W. Wellens	Keys to a successful cleft lip and palate team	Commentary	Interdisciplinary	Cleft lip/palate	
2006	William B. Brinkman	Evaluation of resident communication skills and professionalism: a matter of perspective?	Evaluation	Medicine	Pediatrics	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2006	William J. Swartz	Using gross anatomy to teach and assess professionalism in the first year of medical school	Program development	Medicine	Anatomy	Teamwork/professionalism teaching in the gross anatomy lab; first year of medical school
2007	A.M. Chiesa	An educational process to strengthen primary care nursing practices in São Paulo, Brazil	Program development	Nursing	Family medicine	
2007	Helen Cleak	Preparing health science students for interdisciplinary professional practice	Program Implementation	Interdisciplinary	General	
2007	L. Birch	Obstetric skills drills: evaluation of teaching methods	Evaluation	Interdisciplinary	Obstetrics	
2007	S. Lesinskiene	Use of the HoNOSCA scale in the teamwork of inpatient child psychiatry unit	Evaluation	Interdisciplinary	Psychiatry	Diagnostic scale for use in child psychiatry, not a measure of teamwork
2007	Sarah J. Rudy	Team management training using crisis resource management results in perceived benefits by healthcare workers	Evaluation	Interdisciplinary	General	
2007	Tom W. Reader	Communication skills and error in the intensive care unit	Review	Medicine	Critical care	
2007	V.R. Curran	A framework for integrating interprofessional education curriculum in the health sciences	Evaluation	Interdisciplinary	General	
2007	Vernon R. Curran	Attitudes of health sciences faculty members towards interprofessional teamwork and education	Attitudes	Interdisciplinary	General	Evaluation of interdisciplinary training program

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2008	Andreas Xyrichis	What fosters or prevents interprofessional team working in primary and community care? A literature review	Review	Interdisciplinary	General	Thematic analysis of the literature to identify processes impacting teamwork; barriers to process
2008	Chayan Chakraborti	A systematic review of teamwork training interventions in medical student and resident education	Review	Medicine	General	
2008	Chris Hughes	eMed Teamwork: a self-moderating system to gather peer feedback for developing and assessing teamwork skills	Peer feedback	Medicine	General	
2008	Christopher M. Hicks	Building a simulation-based crisis resource management course for emergency medicine, phase 1: Results from an interdisciplinary needs assessment survey	Program development	Interdisciplinary	Emergency/Trauma	
2008	Dilip R. Patel	Team processes and team care for children with developmental disabilities	Review	Interdisciplinary	Psychiatry	
2008	Eloise Nolan	Teamwork in primary care mental health: a policy analysis	Policy analysis	Medicine	Psychiatry	
2008	Emmanuelle Careau	Assessing interprofessional teamwork in a videoconference-based telerehabilitation setting	Evaluation	Interdisciplinary	Physical medicine and rehabilitation	
2008	Gillian Nisbet	Interprofessional learning for pre-qualification health care students: an outcomes-based evaluation	Evaluation	Interdisciplinary	General	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2008	Guy Haller	Effect of crew resource management training in a multidisciplinary obstetrical setting	Evaluation	Interdisciplinary	Obstetrics	
2008	Haim Berkenstadt	Improving handoff communications in critical care: utilizing simulation-based training toward process improvement in managing patient risk	Evaluation	Nursing	General	
2008	Janet R. Buelow	Building interdisciplinary teamwork among allied health students through live clinical case simulations	Program development	Interdisciplinary	General	
2008	Janine C. Edwards	Promoting regional disaster preparedness among rural hospitals	Quality assurance	Interdisciplinary	General	Disaster preparedness
2008	Jeffrey Damon Dagnone	Interprofessional resuscitation rounds: a teamwork approach to ACLS education	Evaluation	Interdisciplinary	General	
2008	John T. Paige	Implementation of a preoperative briefing protocol improves accuracy of teamwork assessment in the operating room	Evaluation	Interdisciplinary	Surgery	
2008	Marc J. Shapiro	Defining team performance for simulation-based training: methodology, metrics, and opportunities for emergency medicine	Theory	Medicine	Emergency/Trauma	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2008	Niraj L. Sehgal	A multidisciplinary teamwork training program: the triad for optimal patient safety (TOPS) experience	Evaluation	Interdisciplinary	Internal medicine	
2008	Peter J. Pronovost	Improving patient safety in intensive care units in Michigan	Evaluation	Interdisciplinary	Critical care	
2008	Rosemarie Fernandez	Toward a definition of teamwork in emergency medicine	Theory	Medicine	Emergency/Trauma	
2008	Terri E. Weaver	Enhancing multiple disciplinary teamwork	Commentary	Interdisciplinary	Research	
2009	Amy L. Halverson	Surgical team training: the Northwestern Memorial Hospital experience	Attitudes	Interdisciplinary	Surgery	
2009	Andrea Cameron	An introduction to teamwork: findings from an evaluation of an interprofessional education experience for 1000 first-year health science students	Attitudes	Interdisciplinary	General	
2009	Anna R. Gagliardi	Identifying opportunities for quality improvement in surgical site infection prevention	Program development	Medicine	Surgery	
2009	Beatrice J. Kalisch	What does nursing teamwork look like? A qualitative study	Theory	Nursing	General	Qualitative research
2009	Della Freeth	Multidisciplinary obstetric simulated emergency scenarios (MOSES): promoting patient safety in obstetrics with teamwork-focused interprofessional simulations	Evaluation	Interdisciplinary	Obstetrics	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2009	Karen Stead	Teams communicating through STEPPS	Evaluation	Interdisciplinary	General	
2009	Karin Hallin	Active interprofessional education in a patient based setting increases perceived collaborative and professional competence	Evaluation	Interdisciplinary	General	
2009	Leslie W. Hall	Linking health professional learners and health care workers on action-based improvement teams	Quality assurance	Interdisciplinary	General	
2009	Ling Rothrock	Analyses of team performance in a dynamic task environment	Methods	Interdisciplinary	General	Statistical procedure that “takes into account the correlation structure within team members.” Temporal accuracy
2009	Matthew T. Gettman	Use of high fidelity operating room simulation to assess and teach communication, teamwork and laparoscopic skills: initial experience	Evaluation	Medicine	Urology	
2009	Patricia Frakes	Effective teamwork in trauma management	Program Development	Interdisciplinary	General	
2009	Sue Corbet	Teamwork: how does this relate to the operating room practitioner?	Commentary	Medicine	Surgery	
2009	Susan Lerner	Teaching teamwork in medical education	Program development	Interdisciplinary	General	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2009	T. Manser	Teamwork and patient safety in dynamic domains of healthcare: a review of the literature	Review	Interdisciplinary	General	
2009	Tom W. Reader	Developing a team performance framework for the intensive care unit	Theory	Medicine	Critical care	
2010	Brigid M. Gillespie	The impact of organisational and individual factors on team communication in surgery: a qualitative study	Theory	Interdisciplinary	Surgery	Program development
2010	Gudrun Johansson	Multidisciplinary team, working with elderly persons living in the community: a systematic literature review	Review	Interdisciplinary	Geriatrics	
2010	Helen I. Woodward	What have we learned about interventions to reduce medical errors?	Program development	Interdisciplinary	General	
2010	John R. Boulet	Simulation-based assessment in anesthesiology: requirements for practical implementation	Review	Medicine	Anesthesiology	Review of factors supporting successful implementation of simulation-based assessment
2010	Kamal Nagpal	An evaluation of information transfer through the continuum of surgical care: a feasibility study	Quality assurance	Medicine	Surgery	
2010	Katri Hämeen-Anttila	Professional competencies learned through working on a medication education project	Evaluation	Pharmacy	General	Reports on competencies, medical students said they acquired while working on a medication education project

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2010	Lynne A. Donohue	Track, trigger and teamwork: communication of deterioration in acute medical and surgical wards	Quality assurance	Nursing	General	Protocol development
2010	Myrta Rabinowitz	Storytelling effectively translates TeamSTEPPS skills into practice	Evaluation	Nursing	General	Commentary on teaching methods TeamSTEPPS
2010	Sara Evans-Lacko	Facilitators and barriers to implementing clinical care pathways	Theory	Medicine	General	Implementation of care pathways
2011	Aled Jones	Improving teamwork, trust and safety: an ethnographic study of an interprofessional initiative	Evaluation	Interdisciplinary	Geriatrics	Perception of staff regarding improvements in teamwork
2012	A.M. Aboul-Fotouh	Assessment of patient safety culture among healthcare providers at a teaching hospital in Cairo, Egypt	Quality assurance	Interdisciplinary	General	"Assessed healthcare providers' perceptions of patient safety culture within the organization and determined factors that played a role in patient safety culture"
2012	Andreas H. Meier	A surgical simulation curriculum for senior medical students based on TeamSTEPPS	Program development	Medicine	Surgery	
2012	Anna Chang	Transforming Primary Care Training-Patient-Centered Medical Home Entrustable Professional Activities for Internal Medicine Residents	Program development	Medicine	Internal medicine	

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2012	Annemie Vlayen	A nationwide hospital survey on patient safety culture in Belgian hospitals: setting priorities at the launch of a 5-year patient safety plan	Attitudes	Medicine	General	Measure of patient safety culture
2012	Audrey Lyndon	Predictors of likelihood of speaking up about safety concerns in labour and delivery	Attitudes	Medicine	Obstetrics	Study of likelihood of clinicians speaking up about potential harm to patients
2012	Bradley Peckler	Teamwork in the trauma room evaluation of a multimodal team training program	Evaluation	Medicine	Emergency/Trauma	Evaluation of one-day workshop using simulation
2012	Catherine Ménard	Decision-making in oncology: a selected literature review and some recommendations for the future	Evaluation	Medicine	Oncology	Proposal of constructs to measure in evaluating teamwork in oncology
2012	D. Freeth	A methodological study to compare survey-based and observation-based evaluations of organisational and safety cultures and then compare both approaches with markers of the quality of care	Quality assurance	Medicine	Obstetrics	
2012	David J. Klocko	Development, implementation, and short-term effectiveness of an interprofessional education course in a school of health professions	Evaluation	Interdisciplinary	General	Focused on students' understanding of skills needed for interprofessional work; pretest–posttest design

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2012	Deepti Vyas	An interprofessional course using human patient simulation to teach patient safety and teamwork skills	Evaluation	Pharmacy	Pharmacy	Pretest–posttest evaluation of training program with students responding individually to survey items on knowledge, skills, and attitudes
2012	Hanan J. Aboumatar	Republished: development and evaluation of a 3-day patient safety curriculum to advance knowledge, self-efficacy and system thinking among medical students	Program development	Medicine	General	Patient safety curriculum for medical students
2012	Helen A. Scicluna	Clinical capabilities of graduates of an outcomes-based integrated medical program	Evaluation	Medicine	General	Individual "self-perceived" capability. Results state that "Clinical supervisors rated new program graduates highly capable for teamwork, reflective practice, and communication" although the goal of the study was evaluation of an outcomes-based program
2012	Irmajean Bajnok	Building positive relationships in healthcare: evaluation of the teams of interprofessional staff (TIPS) interprofessional education program	Evaluation	Nursing	General	Measure of satisfaction with educational program: "A comprehensive formative and summative evolution revealed that all teams perceid they benefitted from and engaged in successful team development"

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2012	Lukasz M. Mazur	Quantitative assessment of workload and stressors in clinical radiation oncology	Other	Medicine	Oncology	Measure of stressors in oncology
2012	Maja Djukic	NYU3T: teaching, technology, teamwork: a model for interprofessional education scalability and sustainability	Program description	Interdisciplinary	General	
2012	Marc Tumerman	Increasing medical team cohesion and leadership behaviors using a 360° evaluation process	Evaluation	Medicine	Family medicine	Study of the design and implementation of a 360° evaluation project
2012	Margaret Bearman	Learning surgical communication, leadership and teamwork through simulation	Evaluation	Medicine	Surgery	Participant reaction to training provided in course
2012	Nancy C. Elder	Care for patients with chronic nonmalignant pain with and without chronic opioid prescriptions: a report from the Cincinnati Area Research Group (CARinG) network	Outcomes	Medicine	Family medicine	Study of pain medication care for patients in Family Medicine settings
2012	Nicholas R.A. Symons	An observational study of teamwork skills in shift handover	Quality assurance	Medicine	Surgery	
2012	Pamela Turner	Implementation of TeamSTEPPS in the Emergency Department	Evaluation	Interdisciplinary	Emergency/Trauma	
2012	Priscilla Magrath	Paying for performance and the social relations of health care provision: an anthropological perspective	Theory	Medicine	General	Study of pay for performance and social relationships amongst health providers

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2012	Rebecca Lawton	Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review	Review	Medicine	General	Patient safety framework
2012	Reece Hinchcliff	Evaluation of current Australian health service accreditation processes (ACCREDIT-CAP): protocol for a mixed-method research project	Evaluation	Medicine	General	Accreditation processes
2012	Roxanne Tena-Nelson	Reducing potentially preventable hospital transfers: results from a thirty nursing home collaborative	Evaluation	Medicine	Geriatrics	Study of hospital transfers amongst nursing home patients
2012	Susan Brajtman	Toward better care of delirious patients at the end of life: a pilot study of an interprofessional educational intervention	Evaluation	Interdisciplinary	End of life	Training program evaluation specific to competencies associated with end-of-life care
2012	Svin Deneckere	The European quality of care pathways (EQCP) study on the impact of care pathways on interprofessional teamwork in an acute hospital setting: study protocol: for a cluster randomised controlled trial and evaluation of implementation processes	Evaluation	Interdisciplinary	General	Proposal for a “cluster randomized control trial and evaluation of implementation processes”—care pathways

(continued)

(continued)

Publication year	First author	Title	Study type	Profession	Specialty	Comments
2012	Vernon Curran	An approach to integrating interprofessional education in collaborative mental health care	Psychometric	Medicine	Psychiatry	Evaluation of training program with pretest–posttest design. Attitudes concerning interprofessional teamwork measured
2013	Narelle Aram	Intern underperformance is detected more frequently in emergency medicine rotations	Evaluation	Medicine	Emergency/Trauma	Retrospective study of assessment of interns
2013	Nishchay Mehta	Multidisciplinary difficult airway simulation training: two year evaluation and validation of a novel training approach at a district general hospital based in the UK	Evaluation	Medicine	Interdisciplinary	Evaluation of training program using simulation; evaluation included measures of patient outcome regarding airway fatalities

Appendix 2: Complete Reference List for Publications Included in Review

Abdulhadi, N., Al-Shafaee, M. A., Ostenson, C.-G., Vernby, A., & Wahlström, R. (2006). Quality of interaction between primary health-care providers and patients with type 2 diabetes in Muscat, Oman: an observational study. *BMC family practice*, *7*, 72. doi:10.1186/1471-2296-7-72

Aboul-Fotouh, A. M., Ismail, N. A., Ez Elarab, H. S., & Wassif, G. O. (2012). Assessment of patient safety culture among healthcare providers at a teaching hospital in Cairo, Egypt. *Eastern Mediterranean Health Journal*, *18*(4), 372–377.

Aboumatar, H. J., Thompson, D., Wu, A., Dawson, P., Colbert, J., Marsteller, J., et al. (2012). Republished: development and evaluation of a 3-day patient safety curriculum to advance knowledge, self-efficacy and system thinking among medical students. *Postgraduate Medical Journal*, *88*(1043), 545–551. doi:10.1136/postgradmedj-2011-000463rep

Akl, E. A., Bais, A., Rich, E., Izzo, J., Grant, B. J. B., & Schünemann, H. J. (2006). Brief report: Internal medicine residents', attendings', and nurses' perceptions of the night float system. *Journal of General Internal Medicine*, *21*(5), 494–497. doi:10.1111/j.1525-1497.2006.00434.x

Anderson, J. M., Murphy, A. A., Boyle, K. B., Yaeger, K. A., & Halamek, L. P. (2006). Simulating extracorporeal membrane oxygenation emergencies to improve human performance. Part II: assessment of technical and behavioral skills. *Simulation in Healthcare*, *1*(4), 228–232. doi:10.1097/01.SIH.0000243551.01521.74

Aram, N., Brazil, V., Davin, L., & Greenslade, J. (2013). Intern underperformance is detected more frequently in emergency medicine rotations. *Emergency Medicine Australasia: EMA*, *25*(1), 68–74. doi:10.1111/1742-6723.12031

Bajnok, I., Puddester, D., Macdonald, C. J., Archibald, D., & Kuhl, D. (2012). Building Positive Relationships in Healthcare: Evaluation of the Teams of Interprofessional Staff (TIPS) Interprofessional Education Program. *Contemporary Nurse*. doi:10.5172/conu.2012.1495

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2010). Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire. *Quality & Safety in Health Care*, *19*(6). doi:10.1136/qshc.2009.036129

Bearman, M., O'Brien, R., Anthony, A., Civil, I., Flanagan, B., Jolly, B., et al. (2012). Learning surgical communication, leadership and teamwork through simulation. *Journal of Surgical Education*, *69*(2), 201–207. doi:10.1016/j.jsurg.2011.07.014

Bellamy, A., Fiddian, M., & Nixon, J. (2006). Case reviews: promoting shared learning and collaborative practice. *International Journal of Palliative Nursing*, *12*(4), 158–162.

Berkenstadt, H., Haviv, Y., Tuval, A., Shemesh, Y., Megrill, A., Perry, A., et al. (2008). Improving handoff communications in critical care: utilizing simulation-based training toward process improvement in managing patient risk. *Chest*, *134*(1), 158–162. doi:10.1378/chest.08-0914

Birch, L., Jones, N., Doyle, P. M., Green, P., McLaughlin, A., Champney, C., et al. (2007). Obstetric skills drills: evaluation of teaching methods. *Nurse Education Today*, *27*(8), 915–922. doi:10.1016/j.nedt.2007.01.006

Boulet, J. R., & Murray, D. J. (2010). Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology*, *112*(4), 1041–1052.

Braithwaite, J., Westbrook, M., Nugus, P., Greenfield, D., Travaglia, J., Runciman, W., Westbrook, J. (2012). A four-year, systems-wide intervention promoting interprofessional collaboration. *BMC Health Services Research*, *12*. doi:10.1186/1472-6963-12-99

Brajtman, S., Wright, D., Hall, P., Bush, S. H., & Bekele, E. (2012). Toward better care of delirious patients at the end of life: a pilot study of an interprofessional educational intervention. *Journal of Interprofessional Care*, *26*(5), 422–425. doi:10.3109/13561820.2012.694503

Brinkman, W. B., Geraghty, S. R., Lanphear, B. P., Khoury, J. C., Gonzalez del Rey, J. A., DeWitt, T. G., & Britto, M. T. (2006). Evaluation of resident communication skills and professionalism: a matter of perspective? *Pediatrics*, *118* (4), 1371–1379. doi:10.1542/peds.2005-3214

Buelow, J. R., Rathsack, C., Downs, D., Jorgensen, K., Karges, J. R., & Nelson, D. (2008). Building interdisciplinary teamwork among allied health students through live clinical case simulations. *Journal of Allied Health*, *37*(2), e109–123.

Cameron, A., Rennie, S., DiProspero, L., Langlois, S., Wagner, S., Potvin, M., Reeves, S. (2009). An introduction to teamwork: findings from an evaluation of an interprofessional education experience for 1000 first-year health science students. *Journal of Allied Health*, *38*(4), 220–226.

Capella, J., Smith, S., Philp, A., Putnam, T., Gilbert, C., Fry, W., Remine, S. (2010). Teamwork training improves the clinical care of trauma patients. *Journal of Surgical Education*, *67*(6), 439–443. doi:10.1016/j.jsurg.2010.06.006

Careau, E., Vincent, C., & Noreau, L. (2008). Assessing interprofessional teamwork in a videoconference-based telerehabilitation setting. *Journal of Telemedicine and Telecare*, *14*(8), 427–434. doi:10.1258/jtt.2008.080415

Catchpole, K., Mishra, A., Handa, A., & McCulloch, P. (2008). Teamwork and error in the operating room: analysis of skills and roles. *Annals of Surgery*, *247*(4), 699–706. doi:10.1097/SLA.0b013e3181642ec8

Chakraborti, C., Boonyasai, R. T., Wright, S. M., & Kern, D. E. (2008). A systematic review of teamwork training interventions in medical student and resident education. *Journal of General Internal Medicine*, *23*(6), 846–853. doi:10.1007/s11606-008-0600-6

Chang, A., Bowen, J. L., Buranosky, R. A., Frankel, R. M., Ghosh, N., Rosenblum, M. J., et al. (2012). Transforming Primary Care Training-Patient-Centered Medical

Home Entrustable Professional Activities for Internal Medicine Residents. *Journal of general internal medicine*. doi:10.1007/s11606-012-2193-3

Chiesa, A. M., & Fracolli, L. A. (2007). An educational process to strengthen primary care nursing practices in São Paulo, Brazil. *International Nursing Review*, *54*(4), 398–404. doi:10.1111/j.1466-7657.2007.00554.x

Clark, P. G. (2006). What would a theory of interprofessional education look like? Some suggestions for developing a theoretical framework for teamwork training 1. *Journal of Interprofessional Care*, *20*(6), 577–589. doi:10.1080/13561820600916717

Cleak, H., & Williamson, D. (2007). Preparing health science students for interdisciplinary professional practice. *Journal of Allied Health*, *36*(3), 141–149.

Cole, E., & Crichton, N. (2006). The culture of a trauma team in relation to human factors. *Journal of Clinical Nursing*, *15*(10), 1257–1266. doi:10.1111/j.1365-2702.2006.01566.x

Cooper, S., Cant, R., Porter, J., Missen, K., Sparkes, L., McConnell-Henry, T., & Endacott, R. (2012). Managing patient deterioration: assessing teamwork and individual performance. *Emergency Medicine Journal: EMJ*. doi:10.1136/emermed-2012-201312

Cooper, S., Cant, R., Porter, J., Sellick, K., Somers, G., Kinsman, L., & Nestel, D. (2010). Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation*, *81*(4), 446–452. doi:10.1016/j.resuscitation.2009.11.027

Corbet, S. (2009). Teamwork: how does this relate to the operating room practitioner? *Journal of Perioperative Practice*, *19*(9), 278–281.

Curran, V R, & Sharpe, D. (2007). A framework for integrating interprofessional education curriculum in the health sciences. *Education for Health (Abingdon, England)*, *20*(3), 93.

Curran, V. R., Heath, O., Adey, T., Callahan, T., Craig, D., Hearn, T., et al. (2012). An approach to integrating interprofessional education in collaborative mental health care. *Academic Psychiatry*, *36*(2), 91–95. doi:10.1176/appi.ap.10030045

Curran, V. R, Sharpe, D., & Forristall, J. (2007). Attitudes of health sciences faculty members towards interprofessional teamwork and education. *Medical Education*, *41*(9), 892–896. doi:10.1111/j.1365-2923.2007.02823.x

Curtis, J. R., Cook, D. J., Wall, R. J., Angus, D. C., Bion, J., Kacmarek, R., et al. (2006). Intensive care unit quality improvement: a "how-to" guide for the interdisciplinary team. *Critical Care Medicine*, *34*(1), 211–218.

Dagnone, J. D., McGraw, R. C., Pulling, C. A., & Patteson, A. K. (2008). Interprofessional resuscitation rounds: a teamwork approach to ACLS education. *Medical Teacher*, *30*(2), e49–54. doi:10.1080/01421590701769548

Davenport, D. L., Henderson, W. G., Mosca, C. L., Khuri, S. F., & Mentzer, R. M., Jr. (2007). Risk-adjusted morbidity in teaching hospitals correlates with reported levels of communication and collaboration on surgical teams but not with scale measures of teamwork climate, safety climate, or working conditions. *Journal of the American College of Surgeons*, *205*(6), 778–784. doi:10.1016/j.jamcollsurg.2007.07.039

Deneckere, S., Euwema, M., Lodewijckx, C., Panella, M., Sermeus, W., & Vanhaecht, K. (2012). The European quality of care pathways (EQCP) study on the impact of care pathways on interprofessional teamwork in an acute hospital setting: study protocol: for a cluster randomised controlled trial and evaluation of implementation processes. *Implementation Science: IS*, *7*(1), 47. doi:10.1186/1748-5908-7-47

Djukic, M., Fulmer, T., Adams, J. G., Lee, S., & Triola, M. M. (2012). NYU3T: teaching, technology, teamwork: a model for interprofessional education scalability and sustainability. *The Nursing Clinics of North America*, *47*(3), 333–346. doi:10.1016/j.cnur.2012.05.003

Dobson, R. T., Henry, C. J., Taylor, J. G., Zello, G. A., Lachaine, J., Forbes, D. A., & Keegan, D. L. (2006). Interprofessional health care teams: attitudes and environmental factors associated with participation by community pharmacists. *Journal of Interprofessional Care*, *20*(2), 119–132. doi:10.1080/13561820600614031

Donohue, L. A., & Endacott, R. (2010). Track, trigger and teamwork: communication of deterioration in acute medical and surgical wards. *Intensive & Critical Care Nursing*, *26*(1), 10–17. doi:10.1016/j.iccn.2009.10.006

Edwards, J. C., Kang, J., & Silenas, R. (2008). Promoting regional disaster preparedness among rural hospitals. *The Journal of Rural Health*, *24*(3), 321–325. doi:10.1111/j.1748-0361.2008.00176.x

Elder, N. C., Simmons, T., Regan, S., & Gerrety, E. (2012). Care for patients with chronic nonmalignant pain with and without chronic opioid prescriptions: a report from the Cincinnati Area Research Group (CARinG) network. *Journal of the American Board of Family Medicine*, *25*(5), 652–660. doi:10.3122/jabfm.2012.05.120032

Entin, E. B., Lai, F., & Barach, P. (2006). Training teams for the perioperative environment: a research agenda. *Surgical Innovation*, *13*(3), 170–178. doi:10.1177/1553350606294248

Evans-Lacko, S., Jarrett, M., McCrone, P., & Thornicroft, G. (2010). Facilitators and barriers to implementing clinical care pathways. *BMC Health Services Research*, *10*, 182. doi:10.1186/1472-6963-10-182

Fernandez, R., Kozlowski, S. W. J., Shapiro, M. J., & Salas, E. (2008). Toward a definition of teamwork in emergency medicine. *Academic Emergency Medicine*, *15*(11), 1104–1112. doi:10.1111/j.1553-2712.2008.00250.x

Finstuen, K., & Mangelsdorff, A. D. (2006). Executive competencies in healthcare administration: preceptors of the Army-Baylor University Graduate Program. *The Journal of Health Administration Education*, *23*(2), 199–215.

Flabouris, A., Runciman, W. B., & Levings, B. (2006). Incidents during out-of-hospital patient transportation. *Anaesthesia and Intensive Care*, *34*(2), 228–236.

Frakes, P., Neely, I., & Tudoe, R. (2009). Effective teamwork in trauma management. *Emergency Nurse*, *17*(8), 12–17.

Frankel, A., Gardner, R., Maynard, L., & Kelly, A. (2007). Using the Communication and Teamwork Skills (CATS) Assessment to measure health care team performance. *Joint Commission Journal on Quality and Patient safety/Joint Commission Resources*, *33*(9), 549–558.

Freeth, D, Sandall, J., Allan, T., Warburton, F., Berridge, E. J., Mackintosh, N., et al. (2012). A methodological study to compare survey-based and observation-based evaluations of organisational and safety cultures and then compare both approaches with markers of the quality of care. *Health Technology Assessment*, *16*(25), iii–iv, 1–184. doi:10.3310/hta16250

Freeth, D, Ayida, G., Berridge, E. J., Mackintosh, N., Norris, B., Sadler, C., & Strachan, A. (2009). Multidisciplinary obstetric simulated emergency scenarios (MOSES): promoting patient safety in obstetrics with teamwork-focused interprofessional simulations. *The Journal of Continuing Education in the Health Professions*, *29*(2), 98–104. doi:10.1002/chp.20018

Gagliardi, A. R., Eskicioglu, C., McKenzie, M., Fenech, D., Nathens, A., & McLeod, R. (2009). Gerard, S. O., Neary, V., Apuzzo, D., Giles, M. E., & Krinsley, J. (2006). Implementing an intensive glucose management initiative: strategies for success. *Critical Care Nursing Clinics of North America*, *18*(4), 531–543. doi:10.1016/j.ccell.2006.08.004

Gettman, M. T., Pereira, C. W., Lipsky, K., Wilson, T., Arnold, J. J., Leibovich, B. C., et al. (2009). Use of high fidelity operating room simulation to assess and teach communication, teamwork and laparoscopic skills: initial experience. *The Journal of Urology*, *181*(3), 1289–1296. doi:10.1016/j.juro.2008.11.018

Gillespie, B. M., Chaboyer, W., Longbottom, P., & Wallis, M. (2010). The impact of organisational and individual factors on team communication in surgery: a qualitative study. *International Journal of Nursing Studies*, *47*(6), 732–741. doi:10.1016/j.ijnurstu.2009.11.001

Hall, L. W., Headrick, L. A., Cox, K. R., Deane, K., Gay, J. W., & Brandt, J. (2009). Linking health professional learners and health care workers on action-based improvement teams. *Quality Management in Health Care*, *18*(3), 194–201. doi:10.1097/QMH.0b013e3181aea249

Haller, G., Garnerin, P., Morales, M.-A., Pfister, R., Berner, M., Irion, O., et al. (2008). Effect of crew resource management training in a multidisciplinary obstetrical setting. *International Journal for Quality in Health Care*, *20*(4), 254–263. doi:10.1093/intqhc/mzn018

Hallin, K., Kiessling, A., Waldner, A., & Henriksson, P. (2009). Active interprofessional education in a patient based setting increases perceived collaborative and professional competence. *Medical Teacher*, *31*(2), 151–157. doi:10.1080/01421590802216258

Halverson, A. L., Andersson, J. L., Anderson, K., Lombardo, J., Park, C. S., Rademaker, A. W., & Moorman, D. W. (2009). Surgical team training: the Northwestern Memorial Hospital experience. *Archives of Surgery*, *144*(2), 107–112. doi:10.1001/archsurg.2008.545

Hämeen-Anttila, K., Saano, S., & Vainio, K. (2010). Professional competencies learned through working on a medication education project. *American Journal of Pharmaceutical Education*, *74*(6), 110.

Hamilton, N., Freeman, B. D., Woodhouse, J., Ridley, C., Murray, D., & Klingensmith, M. E. (2009). Team behavior during trauma resuscitation: a simulation-based performance assessment. *Journal of Graduate Medical Education*, *1*(2), 253–259. doi:10.4300/JGME-D-09-00046.1

Handler, S. M., Castle, N. G., Studenski, S. A., Perera, S., Fridsma, D. B., Nace, D. A., & Hanlon, J. T. (2006). Patient safety culture assessment in the nursing home. *Quality & Safety in Health Care*, *15*(6), 400–404. doi:10.1136/qshc.2006.018408

Healey, A. N., Undre, S., Sevdalis, N., Koutantji, M., & Vincent, C. A. (2006). The complexity of measuring interprofessional teamwork in the operating theatre. *Journal of Interprofessional Care*, *20*(5), 485–495. doi:10.1080/13561820600937473

Hicks, C. M., Bandiera, G. W., & Denny, C. J. (2008). Building a simulation-based crisis resource management course for emergency medicine, phase 1: Results from an interdisciplinary needs assessment survey. *Academic Emergency Medicine*, *15*(11), 1136–1143. doi:10.1111/j.1553-2712.2008.00185.x

Hinchcliff, R., Greenfield, D., Moldovan, M., Pawsey, M., Mumford, V., Westbrook, J. I., & Braithwaite, J. (2012). Evaluation of current Australian health service accreditation processes (ACCREDIT-CAP): protocol for a mixed-method research project. *BMJ Open*, *2*(4). doi:10.1136/bmjopen-2012-001726

Hsiung, P.-C., Chang, S.-C., & Lin, Y.-Y. (2006). Evaluation of inpatient clinical training in AIDS care. *Journal of the Formosan Medical Association*, *105*(3), 220–228. doi:10.1016/S0929-6646(09)60309-0

Hughes, C., Toohey, S., & Velan, G. (2008). eMed Teamwork: a self-moderating system to gather peer feedback for developing and assessing teamwork skills. *Medical Teacher*, *30*(1), 5–9. doi:10.1080/01421590701758632

Hull, L., Arora, S., Kassab, E., Kneebone, R., & Sevdalis, N. (2011). Assessment of stress and teamwork in the operating room: an exploratory study. *American Journal of Surgery*, *201*(1), 24–30. doi:10.1016/j.amjsurg.2010.07.039

Hutchinson, A., Cooper, K. L., Dean, J. E., McIntosh, A., Patterson, M., Stride, C. B.,et al. (2006). Use of a safety climate questionnaire in UK health care: factor structure, reliability and usability. *Quality & Safety in Health Care*, *15*(5), 347–353. doi:10.1136/qshc.2005.016584

Johansson, G., Eklund, K., & Gosman-Hedström, G. (2010). Multidisciplinary team, working with elderly persons living in the community: a systematic literature review. *Scandinavian Journal of Occupational Therapy*, *17*(2), 101–116. doi:10.1080/11038120902978096

Jones, A., & Jones, D. (2011). Improving teamwork, trust and safety: an ethnographic study of an interprofessional initiative. *Journal of Interprofessional Care*, *25*(3), 175–181. doi:10.3109/13561820.2010.520248

Kaji, A. H., Langford, V., & Lewis, R. J. (2008). Assessing hospital disaster preparedness: a comparison of an on-site survey, directly observed drill performance, and video analysis of teamwork. *Annals of Emergency Medicine*, *52* (3), 195–201, 201.e1–12. doi:10.1016/j.annemergmed.2007.10.026

Kalisch, B. J., Weaver, S. J., & Salas, E. (2009). What does nursing teamwork look like? A qualitative study. *Journal of Nursing Care Quality*, *24*(4), 298–307. doi:10.1097/NCQ.0b013e3181a001c0

Kenaszchuk, C., Reeves, S., Nicholas, D., & Zwarenstein, M. (2010). Validity and reliability of a multiple-group measurement scale for interprofessional collaboration. *BMC Health Services Research*, *10*, 83. doi:10.1186/1472-6963-10-83

King, D. R., Patel, M. B., Feinstein, A. J., Earle, S. A., Topp, R. F., & Proctor, K. G. (2006). Simulation training for a mass casualty incident: two-year experience at the Army Trauma Training Center. *The Journal of Trauma*, *61*(4), 943–948. doi:10.1097/01.ta.0000233670.97515.3a

Kipp, J., McKim, B., Zieber, C., & Neumann, I. (2006). What motivates managers to coordinate the learning experience of interprofessional student teams in service delivery settings? *Healthcare Management Forum*, *19*(2), 42–48.

Klocko, D. J., Hoggatt Krumwiede, K., Olivares-Urueta, M., & Williamson, J. W. (2012). Development, implementation, and short-term effectiveness of an interprofessional education course in a school of health professions. *Journal of Allied Health*, *41*(1), 14–20.

Knapp, C. (2006). Bronson Methodist Hospital: journey to excellence in quality and safety. *Joint Commission Journal on Quality and Patient Safety/Joint Commission Resources*, *32*(10), 556–563.

Körner, M. (2008). Analysis and development of multiprofessional teams in medical rehabilitation. *Psycho-Social Medicine*, *5*, Doc01.

Lamb, D. (2006). Collaboration in practice–assessment of an RAF CCAST. *British Journal of Nursing*, *15*(10), 552–556.

Lawton, R., McEachan, R. R. C., Giles, S. J., Sirriyeh, R., Watt, I. S., & Wright, J. (2012). Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. *BMJ Quality & Safety*, *21*(5), 369–380. doi:10.1136/bmjqs-2011-000443

Lerner, S., Magrane, D., & Friedman, E. (2009). Teaching teamwork in medical education. *The Mount Sinai Journal of Medicine, New York*, *76*(4), 318–329. doi:10.1002/msj.20129

Lesinskiene, S., Senina, J., & Ranceva, N. (2007). Use of the HoNOSCA scale in the teamwork of inpatient child psychiatry unit. *Journal of Psychiatric and Mental Health Nursing*, *14*(8), 727–733. doi:10.1111/j.1365-2850.2007.01160.x

Lurie, S. J., Schultz, S. H., & Lamanna, G. (2011). Assessing teamwork: a reliable five-question survey. *Family Medicine*, *43*(10), 731–734.

Lyndon, A., Sexton, J. B., Simpson, K. R., Rosenstein, A., Lee, K. A., & Wachter, R. M. (2012). Predictors of likelihood of speaking up about safety concerns in labour and delivery. *BMJ Quality & Safety*, *21*(9), 791–799. doi:10.1136/bmjqs-2010-050211

Magrath, P., & Nichter, M. (2012). Paying for performance and the social relations of health care provision: an anthropological perspective. *Social Science & Medicine (1982)*, *75*(10), 1778–1785. doi:10.1016/j.socscimed.2012.07.025

Mann, S., Pratt, S., Gluck, P., Nielsen, P., Risser, D., Greenberg, P., et al. (2006). Assessing quality obstetrical care: development of standardized measures. *Joint Commission Journal on Quality and Patient Safety/Joint Commission Resources*, *32*(9), 497–505.

Manser, T. (2009). Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiologica Scandinavica*, *53*(2), 143–151. doi:10.1111/j.1399-6576.2008.01717.x

Martin, K. S., Elfrink, V. L., Monsen, K. A., & Bowles, K. H. (2006). Introducing standardized terminologies to nurses: Magic wands and other strategies. *Studies in Health Technology and Informatics*, *122*, 596–599.

Mayer, E. K., Winkler, M. H., Aggarwal, R., Karim, O., Ogden, C., Hrouda, D., et al. (2006). Robotic prostatectomy: the first UK experience. *The International Journal of Medical Robotics & Computer Assisted Surgery: MRCAS*, *2*(4), 321–328. doi:10.1002/rcs.113

Mazur, L. M., Mosaly, P. R., Jackson, M., Chang, S. X., Burkhardt, K. D., Adams, R. D., et al. (2012). Quantitative assessment of workload and stressors in clinical radiation oncology. *International Journal of Radiation Oncology, Biology, Physics*, *83*(5), e571–576. doi:10.1016/j.ijrobp.2012.01.063

Mazzocco, K., Petitti, D. B., Fong, K. T., Bonacum, D., Brookey, J., Graham, S., et al. (2009). Surgical team behaviors and patient outcomes. *American Journal of Surgery*, *197*(5), 678–685. doi:10.1016/j.amjsurg.2008.03.002

McCulloch, P., Mishra, A., Handa, A., Dale, T., Hirst, G., & Catchpole, K. (2009). The effects of McNeil, H. P., Hughes, C. S., Toohey, S. M., & Dowton, S. B. (2006). An innovative outcomes-based medical education program built on adult learning principles. *Medical Teacher*, *28*(6), 527–534. doi:10.1080/01421590600834229

Mehta, N., Boynton, C., Boss, L., Morris, H., & Tatla, T. (2013). Multidisciplinary difficult airway simulation training: two year evaluation and validation of a novel training approach at a District General Hospital based in the UK. *European Archives of Oto-Rhino-Laryngology*, *270*(1), 211–217. doi:10.1007/s00405-012-2131-3

Meier, A. H., Boehler, M. L., McDowell, C. M., Schwind, C., Markwell, S., Roberts, N. K., & Sanfey, H. (2012). A surgical simulation curriculum for senior medical students based on TeamSTEPPS. *Archives of Surgery (Chicago, Ill.: 1960)*, *147*(8), 761–766. doi:10.1001/archsurg.2012.1340

Ménard, C., Merckaert, I., Razavi, D., & Libert, Y. (2012). Decision-making in oncology: a selected literature review and some recommendations for the future. *Current Opinion in Oncology*, *24*(4), 381–390. doi:10.1097/CCO.0b013e328354b2f6

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality & Safety in Health Care*, *18*(2), 104–108. doi:10.1136/qshc.2007.024760

Moran, B. J. (2006). Decision-making and technical factors account for the learning curve in complex surgery. *Journal of Public Health*, *28*(4), 375–378. doi:10.1093/pubmed/fdl048

Nagpal, K., Arora, S., Vats, A., Wong, H. W., Sevdalis, N., Vincent, C., & Moorthy, K. (2012). Failures in communication and information transfer across the surgical care pathway: interview study. *BMJ Quality & Safety*. doi:10.1136/bmjqs-2012-000886

Nagpal, K., Vats, A., Ahmed, K., Vincent, C., & Moorthy, K. (2010). An evaluation of information transfer through the continuum of surgical care: a feasibility study. *Annals of Surgery*, *252*(2), 402–407. doi:10.1097/SLA.0b013e3181e986df

Nisbet, G., Hendry, G. D., Rolls, G., & Field, M. J. (2008). Interprofessional learning for pre-qualification health care students: an outcomes-based evaluation. *Journal of Interprofessional Care*, *22*(1), 57–68. doi:10.1080/13561820701722386

Nolan, E., & Hewison, A. (2008). Teamwork in primary care mental health: a policy analysis. *Journal of Nursing Management*, *16*(6), 649–661. doi:10.1111/j.1365-2934.2007.00766.x

O'Leary, K. J., Boudreau, Y. N., Creden, A. J., Slade, M. E., & Williams, M. V. (2012). Assessment of teamwork during structured interdisciplinary rounds on medical units. *Journal of Hospital Medicine*, *7*(9), 679–683. doi:10.1002/jhm.1970

Orchard, C. A., King, G. A., Khalili, H., & Bezzina, M. B. (2012). Assessment of Interprofessional Team Collaboration Scale (AITCS): development and testing of the instrument. *The Journal of Continuing Education in the Health Professions*, *32*(1), 58–67. doi:10.1002/chp.21123

Paige, J. T., Aaron, D. L., Yang, T., Howell, D. S., Hilton, C. W., Cohn, I., Jr, & Chauvin, S. W. (2008). Implementation of a preoperative briefing protocol improves accuracy of teamwork assessment in the operating room. *The American Surgeon*, *74*(9), 817–823.

Parker Oliver, D., & Peck, M. (2006). Inside the interdisciplinary team experiences of hospice social workers. *Journal of Social Work in End-of-life & Palliative Care*, *2*(3), 7–21. doi:10.1300/J457v02n03_03

Patel, D. R., Pratt, H. D., & Patel, N. D. (2008). Team processes and team care for children with developmental disabilities. *Pediatric Clinics of North America*, *55*(6), 1375–1390, ix. doi:10.1016/j.pcl.2008.09.002

Patterson, P. D., Weaver, M. D., Weaver, S. J., Rosen, M. A., Todorova, G., Weingart, L. R., et al. (2012). Measuring Teamwork and Conflict among Emergency Medical Technician Personnel. *Prehospital Emergency Care*, *16*(1), 98–108. doi:10.3109/10903127.2011.616260

Peckler, B., Prewett, M. S., Campbell, T., & Brannick, M. (2012). Teamwork in the trauma room evaluation of a multimodal team training program. *Journal of Emergencies, Trauma, and Shock*, *5*(1), 23–27. doi:10.4103/0974-2700.93106

Pollard, K. C., Miers, M. E., Gilchrist, M., & Sayers, A. (2006). A comparison of interprofessional perceptions and working relationships among health and social care students: the results of a 3-year intervention. *Health & Social Care in the Community*, *14*(6), 541–552. doi:10.1111/j.1365-2524.2006.00642.x

Pronovost, P. J., Berenholtz, S. M., Goeschel, C., Thom, I., Watson, S. R., Holzmueller, C. G., et al. (2008). Improving patient safety in intensive care units in Michigan. *Journal of Critical Care*, *23*(2), 207–221. doi:10.1016/j.jcrc.2007.09.002

Rabinowitz, M., Johnson, L. E., Mazzapica, D., & O'Leary, J. (2010). Storytelling effectively translates TeamSTEPPS skills into practice. *Journal of Continuing Education in Nursing*, *41*(11), 486–487. doi:10.3928/00220124-20101026-03

Reader, T. W., Flin, R., & Cuthbertson, B. H. (2007). Communication skills and error in the intensive care unit. *Current Opinion in Critical Care*, *13*(6), 732–736. doi:10.1097/MCC.0b013e3282f1bb0e

Reader, T. W., Flin, R., Mearns, K., & Cuthbertson, B. H. (2009). Developing a team performance framework for the intensive care unit. *Critical Care Medicine*, *37*(5), 1787–1793. doi:10.1097/CCM.0b013e31819f0451

Rhodes, M., Ashcroft, R., Atun, R. A., Freeman, G. K., & Jamrozik, K. (2006). Teaching evidence-based medicine to undergraduate medical students: a course integrating ethics, audit, management and clinical epidemiology. *Medical Teacher*, *28*(4), 313–317. doi:10.1080/01421590600624604

Rosen, M. A., Weaver, S. J., Lazzara, E. H., Salas, E., Wu, T., Silvestri, S., et al. (2010). Tools for evaluating team performance in simulation-based training. *Journal of Emergencies, Trauma, and Shock*, *3*(4), 353–359. doi:10.4103/0974-2700.70746

Rothrock, L., Cohen, A., Yin, J., Thiruvengada, H., & Nahum-Shani, I. (2009). Analyses of team performance in a dynamic task environment. *Applied Ergonomics*, *40*(4), 699–706. doi:10.1016/j.apergo.2008.06.004

Rudy, S. J., Polomano, R., Murray, W. B., Henry, J., & Marine, R. (2007). Team management training using crisis resource management results in perceived benefits by healthcare workers. *Journal of Continuing Education in Nursing*, *38*(5), 219–226.

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Annals of Surgery*, *255*(4), 804–809. doi:10.1097/SLA.0b013e31824a9a02

Saravanakumar, K., Rao, S. G., & Cooper, G. M. (2006). The challenges of obesity and obstetric anaesthesia. *Current Opinion in Obstetrics & Gynecology*, *18*(6), 631–635. doi:10.1097/GCO.0b013e3280101019

Schraagen, J. M., Schouten, T., Smit, M., Haas, F., van der Beek, D., van de Ven, J., & Barach, P. (2010). Assessing and improving teamwork in cardiac surgery. *Quality & Safety in Health Care*, *19*(6), e29. doi:10.1136/qshc.2009.040105

Scicluna, H. A., Grimm, M. C., O'Sullivan, A. J., Harris, P., Pilotto, L. S., Jones, P. D., & McNeil, H. P. (2012). Clinical capabilities of graduates of an outcomes-based integrated medical program. *BMC Medical Education*, *12*, 23. doi:10.1186/1472-6920-12-23

Scott-Cawiezell, J., & Vogelsmeier, A. (2006). Nursing home safety: a review of the literature. *Annual Review of Nursing Research*, *24*, 179–215.

Sehgal, N. L., Fox, M., Vidyarthi, A. R., Sharpe, B. A., Gearhart, S., Bookwalter, T., et al. (2008). A multidisciplinary teamwork training program: the Triad for Optimal Patient Safety (TOPS) experience. *Journal of General Internal Medicine*, *23*(12), 2053–2057. doi:10.1007/s11606-008-0793-8

Sevdalis, N., Lyons, M., Healey, A. N., Undre, S., Darzi, A., & Vincent, C. A. (2009). Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Annals of Surgery*, *249*(6), 1047–1051. doi:10.1097/SLA.0b013e3181a50220

Sexton, J. B., Makary, M. A., Tersigni, A. R., Pryor, D., Hendrich, A., Thomas, E. J., et al. (2006). Teamwork in the operating room: frontline perspectives among hospitals and operating room personnel. *Anesthesiology*, *105*(5), 877–884.

Shapiro, M. J., Gardner, R., Godwin, S. A., Jay, G. D., Lindquist, D. G., Salisbury, M. L., & Salas, E. (2008). Defining team performance for simulation-based training: methodology, metrics, and opportunities for emergency medicine. *Academic Emergency Medicine*, *15*(11), 1088–1097. doi:10.1111/j.1553-2712.2008.00251.x

Sharma, B., Mishra, A., Aggarwal, R., & Grantcharov, T. P. (2011). Non-technical skills assessment in surgery. *Surgical Oncology*, *20*(3), 169–177. doi:10.1016/j.suronc.2010.10.001

Sharp, M. (2006). Enhancing interdisciplinary collaboration in primary health care. *Canadian Journal of Dietetic Practice and Research*, *Suppl*, S4–8.

Simpson, K. R., James, D. C., & Knox, G. E. (2006). Nurse-physician communication during labor and birth: implications for patient safety. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, *35*(4), 547–556. doi:10.1111/j.1552-6909.2006.00075.x

Stead, K., Kumar, S., Schultz, T. J., Tiver, S., Pirone, C. J., Adams, R. J., & Wareham, C. A. (2009). Teams communicating through STEPPS. *The Medical Journal of Australia*, *190*(11 Suppl), S128–132.

Swartz, W. J. (2006). Using gross anatomy to teach and assess professionalism in the first year of medical school. *Clinical Anatomy*, *19*(5), 437–441. doi:10.1002/ca.20331

Symons, N. R. A., Wong, H. W. L., Manser, T., Sevdalis, N., Vincent, C. A., & Moorthy, K. (2012). An observational study of teamwork skills in shift handover. *International Journal of Surgery*. doi:10.1016/j.ijsu.2012.05.010

Takayesu, J. K., Farrell, S. E., Evans, A. J., Sullivan, J. E., Pawlowski, J. B., & Gordon, J. A. (2006). How do clinical clerkship students experience simulator-based teaching? A qualitative analysis. *Simulation in Healthcare*, *1*(4), 215–219. doi:10.1097/01.SIH.0000245787.40980.89

Taylor, C., Brown, K., Lamb, B., Harris, J., Sevdalis, N., & Green, J. S. A. (2012). Developing and testing TEAM (Team Evaluation and Assessment Measure), a self-assessment tool to improve cancer multidisciplinary teamwork. *Annals of Surgical Oncology*, *19*(13), 4019–4027. doi:10.1245/s10434-012-2493-1

Tena-Nelson, R., Santos, K., Weingast, E., Amrhein, S., Ouslander, J., & Boockvar, K. (2012). Reducing potentially preventable hospital transfers: results from a thirty nursing home collaborative. *Journal of the American Medical Directors Association*, *13*(7), 651–656. doi:10.1016/j.jamda.2012.06.011

Thomas, E. J., Sexton, J. B., Lasky, R. E., Helmreich, R. L., Crandell, D. S., & Tyson, J. (2006). Teamwork and quality during neonatal care in the delivery room. *Journal of Perinatology*, *26*(3), 163–169. doi:10.1038/sj.jp.7211451

Tumerman, M., & Carlson, L. M. H. (2012). Increasing medical team cohesion and leadership behaviors using a 360-degree evaluation process. *WMJ: Official Publication of the State Medical Society of Wisconsin*, *111*(1), 33–37.

Turner, P. (2012). Implementation of TeamSTEPPS in the Emergency Department. *Critical Care Nursing Quarterly*, *35*(3), 208–212. doi:10.1097/CNQ.0b013e3182542c6c

Undre, S., Healey, A. N., Darzi, A., & Vincent, C. A. (2006). Observational assessment of surgical teamwork: a feasibility study. *World Journal of Surgery*, *30*(10), 1774–1783. doi:10.1007/s00268-005-0488-9

Undre, S., Koutantji, M., Sevdalis, N., Gautama, S., Selvapatt, N., Williams, S., et al. (2007). Multidisciplinary crisis simulations: the way forward for training surgical teams. *World Journal of Surgery*, *31*(9), 1843–1853. doi:10.1007/s00268-007-9128-x

Varkey, P., Gupta, P., Arnold, J. J., & Torsher, L. C. (2009). An innovative team collaboration assessment tool for a quality improvement curriculum. *American Journal of Medical Quality*, *24*(1), 6–11. doi:10.1177/1062860608326420

Vlayen, A., Hellings, J., Claes, N., Peleman, H., & Schrooten, W. (2012). A nationwide hospital survey on patient safety culture in Belgian hospitals: setting priorities at the launch of a 5-year patient safety plan. *BMJ Quality & Safety*, *21*(9), 760–767. doi:10.1136/bmjqs-2011-051607

Vyas, D., McCulloh, R., Dyer, C., Gregory, G., & Higbee, D. (2012). An interprofessional course using human patient simulation to teach patient safety and teamwork skills. *American Journal of Pharmaceutical Education*, *76*(4), 71. doi:10.5688/ajpe76471

Weaver, T. E. (2008). Enhancing multiple disciplinary teamwork. *Nursing Outlook*, *56*(3), 108–114.e2. doi:10.1016/j.outlook.2008.03.013

Wellens, W., & Vander Poorten, V. (2006). Keys to a successful cleft lip and palate team. *B-ENT*, *2 Suppl 4*, 3–10.

Wholey, D. R., Zhu, X., Knoke, D., Shah, P., Zellmer-Bruhn, M., & Witheridge, T. F. (2012). The Teamwork in Assertive Community Treatment (TACT) Scale: Development and Validation. *Psychiatric Services*, *63*(11), 1108–1117. doi:10.1176/appi.ps.201100338

Woodward, H. I., Mytton, O. T., Lemer, C., Yardley, I.E., Ellis, B. M., Rutter, P. D., Wu, A. W. (2010). What have we learned about interventions to reduce medical errors? *Annual Review of Public Health*, *31*, 479–497 1 p following 497. doi:10.1146/annurev.publhealth.012809.103544

Wright, M. C., Phillips-Bute, B. G., Petrusa, E. R., Griffin, K. L., Hobbs, G. W., & Taekman, J. M. (2009). Assessing teamwork in medical education and practice:

relating behavioural teamwork ratings and clinical performance. *Medical Teacher*, *31*(1), 30–38. doi:10.1080/01421590802070853

Xyrichis, A., & Lowton, K. (2008). What fosters or prevents interprofessional team working in primary and community care? A literature review. *International Journal of Nursing Studies*, *45*(1), 140–153. doi:10.1016/j.ijnurstu.2007.01.015

Yildirim, A., Akinci, F., Ates, M., Ross, T., Issever, H., Isci, E., & Selimen, D. (2006). Turkish version of the Jefferson Scale of Attitudes Toward Physician-Nurse Collaboration: a preliminary study. *Contemporary Nurse*, *23*(1), 38–45. doi:10.5555/conu.2006.23.1.38

Yule, S, Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery*, *139*(2), 140–149. doi:10.1016/j.surg.2005.06.017

Yule, S, Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Medical Education*, *40*(11), 1098–1104. doi:10.1111/j.1365-2929.2006.02610.x

Yule, Steven, Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World Journal of Surgery*, *32*(4), 548–556. doi:10.1007/s00268-007-9320-z

References

Aboumatar, H. J., Thompson, D., Wu, A., Dawson, P., Colbert, J., Marsteller, J., & Pronovost, P. (2012). Republished: development and evaluation of a 3-day patient safety curriculum to advance knowledge, self-efficacy and system thinking among medical students. *Postgraduate Medical Journal, 88*(1043), 545–551. doi:10.1136/postgradmedj-2011-000463rep

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2010). Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire. *Quality & Safety in Health Care, 19*(6), e49. doi:10.1136/qshc.2009.036129

Baker, D. P., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research, 41*(4 Pt 2), 1576–1598. doi:10.1111/j.1475-6773.2006.00566.x

Baker, D. P., & Salas, E. (1992). Principles for measuring teamwork skills. *Human Factors, 34*(4), 469–475. doi:10.1177/001872089203400408

Capella, J., Smith, S., Philp, A., Putnam, T., Gilbert, C., Fry, W., & Remine, S. (2010). Teamwork training improves the clinical care of trauma patients. *Journal of Surgical Education, 67*(6), 439–443. doi:10.1016/j.jsurg.2010.06.006

Driskell, J. E., & Salas, E. (1992). Collective behavior and team performance. *Human Factors, 34* (3), 277–288. doi:10.1177/001872089203400303

Eva, K. W., Cunnington, J. P. W., Reiter, H. I., Keane, D. R., & Norman, G. R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Advances in Health Sciences Education: Theory and Practice, 9*(3), 211–224.

Eva, K. W., & Regehr, G. (2010). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3), 311–329. doi:10.1007/s10459-010-9263-2

Healey, A. N., Undre, S., Sevdalis, N., Koutantji, M., & Vincent, C. A. (2006). The complexity of measuring interprofessional teamwork in the operating theatre. *Journal of Interprofessional Care, 20*(5), 485–495. doi:10.1080/13561820600937473

Kenaszchuk, C., Reeves, S., Nicholas, D., & Zwarenstein, M. (2010). Validity and reliability of a multiple-group measurement scale for interprofessional collaboration. *BMC Health Services Research, 10*, 83. doi:10.1186/1472-6963-10-83

Kendall, D., & Salas, E. (2004). Measuring Team Performance: Review of Current Methods and Consideration of Future Needs. In *Advances in Human Performance and Cognitive Engineering Research* (Vol. 5, pp. 307–326). Elsevier. Retrieved from http://www.emeraldinsight.com/10.1016/S1479-3601(04)05006-4

Lawton, R., McEachan, R. R. C., Giles, S. J., Sirriyeh, R., Watt, I. S., & Wright, J. (2012). Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. *BMJ Quality & Safety, 21*(5), 369–380. doi:10.1136/bmjqs-2011-000443

Lerner, S., Magrane, D., & Friedman, E. (2009). Teaching teamwork in medical education. *The Mount Sinai Journal of Medicine, New York, 76*(4), 318–329. doi:10.1002/msj.20129

Lurie, S. J., Schultz, S. H., & Lamanna, G. (2011). Assessing teamwork: a reliable five-question survey. *Family Medicine, 43*(10), 731–734.

Mazzocco, K., Petitti, D. B., Fong, K. T., Bonacum, D., Brookey, J., Graham, S., & Thomas, E. J. (2009). Surgical team behaviors and patient outcomes. *American Journal of Surgery, 197*(5), 678–685. doi:10.1016/j.amjsurg.2008.03.002

Morgan, P. J., Pittini, R., Regehr, G., Marrs, C., & Haley, M. F. (2007). Evaluating teamwork in a simulated obstetric environment. *Anesthesiology, 106*(5), 907–915. doi:10.1097/01.anes.0000265149.94190.04

Murray, D., & Enarson, C. (2007). Communication and teamwork. *Anesthesiology, 106*(5), 895–896. doi:10.1097/01.anes.0000265145.40825.ac

Nagpal, K., Arora, S., Vats, A., Wong, H. W., Sevdalis, N., Vincent, C., & Moorthy, K. (2012). Failures in communication and information transfer across the surgical care pathway: interview study. *BMJ Quality & Safety,*. doi:10.1136/bmjqs-2012-000886

O'Leary, K. J., Boudreau, Y. N., Creden, A. J., Slade, M. E., & Williams, M. V. (2012). Assessment of teamwork during structured interdisciplinary rounds on medical units. *Journal of hospital medicine: an official publication of the Society of Hospital Medicine, 7*(9), 679–683. doi:10.1002/jhm.1970

Orchard, C. A., King, G. A., Khalili, H., & Bezzina, M. B. (2012). Assessment of interprofessional team collaboration scale (AITCS): development and testing of the instrument. *The Journal of continuing education in the health professions, 32*(1), 58–67. doi:10.1002/chp.21123

Patterson, P. D., Weaver, M. D., Weaver, S. J., Rosen, M. A., Todorova, G., Weingart, L. R., & Salas, E. (2012). Measuring teamwork and conflict among emergency medical technician personnel. *Prehospital Emergency Care, 16*(1), 98–108. doi:10.3109/10903127.2011.616260

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Annals of Surgery, 255*(4), 804–809. doi:10.1097/SLA.0b013e31824a9a02

Salas, Eduardo, DiazGranados, D., Weaver, S. J., & King, H. (2008). Does team training work? Principles for health care. *Academic Emergency Medicine, 15*(11), 1002–1009. doi:10.1111/j.1553-2712.2008.00254.x

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork? *Small Group Research, 36*(5), 555–599. doi:10.1177/1046496405277134

Tumerman, M., & Carlson, L. M. H. (2012). Increasing medical team cohesion and leadership behaviors using a 360-degree evaluation process. *WMJ: official publication of the State Medical Society of Wisconsin, 111*(1), 33–37.

Varkey, P., Gupta, P., Arnold, J. J., & Torsher, L. C. (2009). An innovative team collaboration assessment tool for a quality improvement curriculum. *American journal of medical quality: the official journal of the American College of Medical Quality, 24*(1), 6–11. doi:10.1177/1062860608326420

Vyas, D., McCulloh, R., Dyer, C., Gregory, G., & Higbee, D. (2012). An interprofessional course using human patient simulation to teach patient safety and teamwork skills. *American journal of pharmaceutical education, 76*(4), 71. doi:10.5688/ajpe76471

Chapter 15
Developing and Assessing Teams Working Collaboratively Across Professions

Anne McKee

Abstract Team based working is vital to the delivery of high quality care. Most medical and health professionals understand this. However developing and assessing effective collaborative practice remains troublesome. A study to develop and assess multi-professional learning organisations in primary care in the United Kingdom (U.K.) provided an opportunity for detailed examination of how to enable this form of team based learning and assessment in primary care clinical settings. The key findings of the study identify the core competencies needed to enable engagement in cross professional and interdisciplinary learning and assessment and argue for a re-thinking of assessment approaches to collaborative team working. Both policy and practice contexts had a significant impact on engagement in learning and assessment. It is argued that an approach to assessment is needed that takes context into account and re-emphasizes assessment focused on learning and improvement of collaborative working practices.

Takeaways

- Team based working is vital to the delivery of high quality care. Most medical and health professionals understand this. However developing and assessing effective collaborative practice remains troublesome.
- A study to develop and assess multi-professional learning organizations in primary care in the United Kingdom (U.K.) provided an opportunity for detailed examination of how to enable this form of team learning and assessment in primary care clinical settings.
- The key findings of the study identify the core competencies needed to enable engagement in cross-professional and interdisciplinary learning and assessment and argue for a re-thinking of assessment approaches to collaborative team working. Both policy and practice contexts had a significant impact on engagement in learning and assessment. It is argued

A. McKee (✉)
Division of Medical Education, King's College London, London, UK
e-mail: anne.mckee@kcl.ac.uk

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_15

that an approach to assessment is needed that takes contexts into account and re-emphasizes assessment focused on learning and improvement of collaborative working practices.

15.1 Introduction

This chapter examines several challenges when conducting work-based assessments of team learning in primary care. Using action research as a method for developing and assessing teams working collaboratively, this study probes the effects of how the contexts of practice influence both engagement in learning and assessment of learning. This brings into sharp relief the social situatedness of both. Understanding how to improve team-based assessment of learning is becoming increasingly urgent in primary care in the United Kingdom (U.K.) following national changes in funding for continuing professional development. The U.K. case explored here, addresses a central theme of this book, which is to illuminate assessment issues, developing a better understanding of their complexity, and the implications of this for assessment of learning in interprofessional and cross-professional practice.

Within professional education, the development of performance assessment predominantly focuses on performance of *individuals* more than *teams.* Assessment for licensure, postgraduate specialization, and revalidation creates an educational emphasis that is essentially oriented toward individuals, within their own professional group. However, providing care and healthcare practice requires team working. This is recognized by the General Medical Council and forms part of their national standards framework.

Are there limitations to this recurring focus on the assessment of individual performance? This question is examined within an initiative involving primary healthcare teams in the United Kingdom. Like many practice contexts, primary care requires teams to work collaboratively. Those responsible for training primary care practitioners in East Anglia[1] commissioned an action research project to develop learning organizations. They believed this could increase capacity for training placements. The research and development project illuminated contemporary realities of working in primary care. Policy assumptions and practice realities were disconnected. Findings from the study emphasized the need to reconsider how learning is a process embedded in social relationships, with practices, purposes, and implications for team and organizational assessment.

The study identifies: (1) The impact of policy on practice, learning, assessment, and accountability when practicalities matter. (2) Complexities of administering

[1] Primary care settings in this project refer to general practice clinical settings.

comparable assessments of work-based learning when stakeholders[2] and primary care professionals interpret project purposes and outcomes differently. (3) Challenges when developing *practitioner-conducted assessments of learning* arising from everyday practice where heavy workloads couple with high external demands.

Key characteristics of the workplace of primary care were identified and these suggested the need to rethink assessment-for-improvement that is team and organizationally-based. Approaches to developing team and organizational assessments are proposed.

15.2 Why Develop Learning Organisations?

The East of England deanery, the body responsible for training primary care practitioners, faced a pressing problem. They needed to increase capacity within the region for training placements. Taking the initiative, they decided to develop learning organizations as a means of creating suitable learning environments for a range of primary care health professionals in training. The deanery commissioned an action research project in the East of England (2009–2011) to develop multi-professional learning organizations in primary care, called: "*Developing Multi-Professional Learning Organizations in Primary Care* (*MPLO*)."

Action research is an established methodology valued because it involves a process of connecting research to development. It is undertaken with the express intention of informing and improving what people do. Kurt Lewin (1890–1947) is often thought to be the founder of action research. His aim was to direct research toward solving social problems. The "action" orientation of action research is its solution or developmental focus.

The aim of this project was to help primary care teams based in general practice to develop a productive culture of work-based learning and reflective practice, thus enriching the learning environment within practices. A pilot study undertaken in seven general practices between January and April 2009 evidenced enthusiasm for such a project, as an opportunity to develop work-based learning that would support the development of primary care teams in changing contexts.

15.3 Why Did Primary Care Practices Express Enthusiasm for the Project?

Most of the practices who took part in the pilot hoped that the project would enable them to respond and develop positively in a service context they described as relentlessly high volume and high demand. Some talked of feeling beleaguered and

[2] A stakeholder is a person, group or organization who has an interest in or is responsible for an initiative.

concerned that they were not engaging with challenges as creatively as they felt capable of doing. Others, already developing as learning organizations, hoped for support and sharing with other practices.

However, even the enthusiasts expressed caution. The semi-structured pilot interviews documented some of the work pressures they faced. These included practical, resource, and staff development considerations that needed to be addressed if practices were to become or develop further as learning organizations.

Some practical challenges to becoming a learning organization included the lack of:

- *Physical space for accommodating trainees*. A private space in which to have learning conversations, particularly sensitive ones, was either nonexistent or in much demand within practices.
- *Physical space for group or team meetings*. Places for group or team meetings were also in short supply but in high demand.
- *Time to think, learn, and reflect*. Busy, demand-led workloads made engaging in learning and teaching challenging.
- *Motivation*. Doctors raised concerns about burn-out, exhaustion, and cynicism —their own and others.
- *Resources for clinical cover to enable general practitioners to attend external events, and resources to free members of the primary care team to engage in staff development* inside and outside the practice.

Those practices that were enthusiastic about developing as a learning organization saw this as an opportunity to thrive and survive demanding service pressures and to have their challenges understood and addressed. The East Anglian Deanery deliberated on the pilot and decided to fund Phase 2 of the project. However, in the early stages of Phase 2, the climate in which the project operated changed significantly. A new health policy was introduced called *Liberating the NHS: Developing the Healthcare Workforce: From Design to Delivery* (Department of Health 2010). This policy reemphasized the important role of continuing professional development in improving the healthcare system, but within a structure that would determine and apply national learning outcomes. This policy also initiated a review of the organizational structures in which professional education was provided. The deanery, as a provider of health professional education, was subject to organizational restructuring.

This policy had consequences for both the deanery and primary care practices because it created an uncertain policy context. The speed and scope of the proposed changes concerned both practitioners and professional bodies. Their lobbying resulted in the creation of a national consultation process.

Liberating the NHS: Developing the Healthcare Workforce: From Design to Delivery proved to be a radical policy initiative that had a significant impact on both the brief and progress of the MPLO project. This policy raised questions about how work-based learning initiatives were valued within a reorganized continuing professional development landscape.

15.4 Policy Context: A Closer Look

Learning organizations have been part of the discourse within medical education for some time. They form part of a response to policies associated with continuing professional development in primary care. These policies attempt to align personal and organizational development. For example in 1998, the Chief Medical Officers' report, *A Review of Continuing Professional Development* articulated a refocusing of the role of work-based learning in professional development (Department of Health 1998). The report encouraged the development of Professional Practice Development Plans (PPDP). These linked personal professional development with practice development. This policy proved both challenging and problematic because pressures to comply with external agendas competed with needs of individual practices (McKee and Watts 2003). The policy pursuit of meaningful engagement with professional development continued when on 31 March 2004, the Postgraduate Educational Allowance (PGEA) ended. PGEA was an established system offering credit to individual General Practitioners who attended approved learning events. PGEA, was designed to incentivize attendance at approved events. Performance assessment was not in its remit. Removing PGEA was a quiet but significant departure from gently encouraging engagement with continuing professional development.

Since 1998, national frameworks and mechanisms have established an infrastructure and tools with which to externally manage and review primary care organizational systems and those who work within them (Department of Health 1998). Examples of these mechanisms include: annual individual professional appraisal and organizational review through the Quality Outcomes Framework (QOF) (hscic 2016). In the quirkiness of the National Health Service, this policy involved the erosion of long-held autonomy in primary care that was predominantly clinical and led by general practitioners. The evolving policy trend was that *learning and performance management agendas have been coupled in new kinds of alliances.*

More recently in *Liberating the NHS: Developing the Healthcare Workforce: From Design to Delivery* (Department of Health 2012), policy makers argued that education and training needed to be more flexible and responsive to changing healthcare demands and new patterns of health care. This policy had the following results:

- Employers were given a role in educational commissioning and governance.
- New organizational structures were to be put in place, reflecting a further strengthening of public accountability.

From an educational perspective these changes imply the need for:

- An extended stakeholder curriculum, with
- Learning outcomes set nationally,
- Education and training provided locally to meet both national education outcomes and local needs,
- Within a top-down accountability structure.

This policy generated uncertainty, not just because of the scale of proposed changes but also because of the planned speed with which those changes were to be introduced. Deaneries responsible for the training of primary care practitioners faced a process of restructuring and downsizing. During this project, a Deanery Commissioner of the MPLO project and a number of administrative staff lost their posts.

15.5 Policy Impacts on the MPLO Project

The MPLO project team encountered a *climate of demoralization*, that is low morale, within the deanery and primary care practices. This involved a reduction in their responsiveness to project activities.

Within the deanery, redundancies meant stretched resources and new challenges in supporting the project. Communication suffered with very delayed responses to email and phone messages. At times these delays took two or more months, the timely financial administration of the project also suffered. As a result, one member of the project team withdrew to work elsewhere.

Within participating primary care practices, it seemed that evolving performance management and educational agendas had created a *climate of caution*, which influenced how practice members engaged with the project.

For the MPLO project team, policy effects had created unpredictable challenges in engaging with the brief that outlined the project. The director and principal investigator reconsidered how to undertake work-based development initiatives and how to determine their effectiveness.

15.6 Theoretical Perspectives

The main theoretical perspectives deployed in the study were *Learning Organizations* and *Thinking Organizations.* Learning organizations are prized because of their perceived ability to enable responsive and flexible approaches to work-based learning. Learning organizations provide an approach to developing people and systems in change contexts. Senge's definition of a learning organization is quoted here to claim that such organizations have powerful attributes that will ensure success and quality.

You can gradually evolve a new type of organization. It will be able to deal with the problems and opportunities of today, and invest in its capacity to embrace tomorrow, because its members are continually focusing on enhancing and expanding their collective awareness and abilities. You can create, in other words, an organization which can learn (Senge 1994).

Table 15.1 Learning organizations: key themes

Characteristics	• Flatter, team-based structures • Values—prioritizing learning • Values—prioritizing empowerment for change
Becoming a learning practice	• Individual and organizational learning begins a process leading toward a learning culture • Routines need to be established that create a supportive, systematic approach to learning which, in turn, creates conditions that make learning integral to what a practice does
Core conditions	• Strong, visionary leadership that: 1. supports and develops others 2. asks challenging questions 3. is willing to learn 4. sees possibilities 5. makes things happen 6. facilitates learning environments. • Involvement and empowerment of staff where changes grow from the willing participation of all. • Setting aside times and places for reflection (Rushmer et al. 2004c)

What are learning organizations and how do they "excel"? Argyris and Schön (1978) identified the distinguishing levels of learning that characterized learning organizations. They argued that double-loop learning involves a shift from the application of routinized or established practice to the ability to think critically and modify practice in response to the uncertain, unpredictable, and particular challenges that practitioners encounter. At this level, practitioners and learners can question assumptions. They process both their formal and informal learning to develop new responses or practices.

How are learning organizations described and understood, within the educational literature on primary care? In a series of three papers based on empirical research into learning organizations in healthcare settings, Rushmer et al. (2004a, b, c), examine characteristics and conditions for becoming a learning organization in primary care. Key themes from the papers are summarized in Table 15.1.

These characteristics provided a set of assessment criteria to help identify learning organizations and the maturity of their development.

15.6.1 Thinking Organizations

From the literature on organizational development, the conditions for *thinking organizations* shored up the characteristics of a learning organization described by Kelly et al., but further elaborated the required conditions in terms of values, culture, and practices. In order for learning to take place, certain conditions need to be in place. Foremost among these are the conditions that make a thinking environment possible. Nancy Kline has developed her idea of the thinking environment

Table 15.2 Thinking environment components

Attention	Respectful listening and thoughtful attention have a powerful effect on others
Equality	Valuing everyone equally, ensuring equal time and attention, respecting boundaries and agreements to enable thinking to be articulated
Appreciation	Focusing on the positive provides a balanced view that is not only seeking to identify problems but also recognizes the good in every situation
Ease	People need to feel relaxed and comfortable if they are to think clearly. Freedom from pressure and a sense of urgency will contribute to a creative, thoughtful environment
Encouragement	Thoughts and their thinkers should be encouraged on their own merits in order to reduce the sense of competition which can stifle creativity and give courage to explore the frontiers of the thought

over many years. In her book, *Time to Think* (1999), she sets out ten components that are the foundation of the thinking environment. The five described below proved valuable when analyzing the data from participating MPLO primary care practices and identifying core competencies and conditions needed to function as a learning organization (Table 15.2).

These theoretical perspectives informed our thinking about what a learning organization was or might be, and the processes that would best support their development and assessment within a work-based learning approach.

The *Royal College of General Practitioners* (RCGP) has a module within their *Quality Practice Award* on learning organizations. This seeks to involve practices in established good practice, such as ensuring appropriate qualifications to undertake roles, annual appraisals for nonclinical staff, significant event reviews, patient involvement, and a commitment to working effectively together (Royal College of General Practitioners Quality Practice Award 2012).

The action research-based approach of the *Developing Multi-Professional Learning Organizations in Primary Care* project was different but complementary to that of the RCGP.

The project team focused on the starting point and needs of individual practices.

The project team approach involved *benchmarking* the starting point of practices in relation to their learning needs and concerns. Through facilitation, participating practices were supported to address their needs and concerns using an action research inquiry process. Collaborative working, identifying problems or issues, and co-creating action to address those issues, characterize action research methodology. Developing multiprofessional learning teams involved enabling:

- Shifts in organizational systems and cultures,
- Concepts and practices of leadership that are responsive to implementing change,
- Understanding concepts and skills on how to facilitate the learning process, and
- A culture of reflective practice.

The action research pedagogical approach to support work-based learning was process oriented. It acknowledged the need for individual, group, and team learning. Focusing on the starting point and needs of individual practices, the proposed curriculum was work-based, work-focused, and sensitive to contexts and diverse needs.

15.7 Methodology

The methodology of action research supports evidence-based and reflective practice to enable the learning processes necessary for shifts in values, culture, and practices, associated with learning organizations and thinking environments. Action research is undertaken with the express intention of informing and improving what people do (Reason and Bradbury 2001). Development is at its heart. It is often used to address practical problems or pressing issues in a process of inquiry that embeds cyclical development. Typically, this involves framing a question or concern, designing, and conducting an investigation, identifying and implementing a change, reviewing that change, and implementing any necessary further changes. Action research reveals the complexities of a situation and the social practices or social behaviors associated with it. *It takes account of contexts. This was to prove essential* (Fig. 15.1).

As previously described there were two phases to the project:

- Phase 1: January–April 2009
- Phase 2: September 2010–June 2011

In Phase 1 of the project (January–April 2009), we conducted a pilot test of the feasibility of developing learning organizations using a development and research strategy. Seven primary care practices participated in the pilot. They were

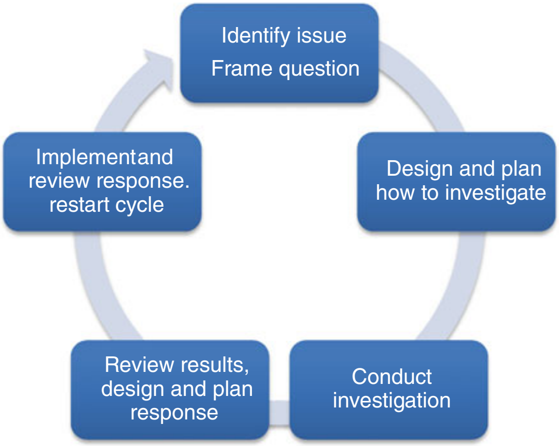

Identify issue
Frame question
Design and plan how to investigate
Conduct investigation
Review results, design and plan response
Implementand review response. restart cycle

Fig. 15.1 The typical action research cycle

recognized and regulated training practices within the deanery. Training status is a mark of quality in organizational systems, teaching, and by proxy, provision of care. The pilot practices were fertile sites for development.

From these practices, seven doctors, three nurses, and six practice managers took part in either individual or group interviews. This sample was smaller than hoped for, due to delays in accessing accurate contact details, and securing time to meet practice members. The semi-structured interview schedule used in pilot interviews documented some of the work pressures primary care practitioners faced. Interviews identified practical, resource, and staff development issues that needed to be addressed if practices were to become or develop further as learning organizations. They proved to be honest and insightful descriptors of conditions within primary care practices that would prove challenging for both practices and the project team.

15.7.1 Phase 2: Original Design and Empirical Adjustments

The project was established as a research and development project involving first- and second-order action research (Elliott 1991). The first-order action research involves the activities of those participating in the MPLO project as they use the action research cycle. This cycle is a research-based approach to development. The second-order action research involved the MPLO project team as they conducted research into what was happening and with what effects. Within the MPLO project, the principal investigator facilitated primary care practices in their development activities and collected data. Data were analyzed by team members who did not have that facilitating role.

MPLO facilitation of primary care practices. Participating practices were supported in identifying their learning needs, formulating a development project, and implementing that project. Key to this process was helping primary care practices develop reflective and evidence-based approaches to practice as they moved through cycles of action, review, and reflection.

Methods to support this process included:

- Reflective logs,
- Recorded discussions,
- A learning plan,
- Self-assessment questionnaires, and
- A final review or report of learning in portfolio form.

The first-order action research involved the design and implementation of practice development plans. This included development and review of primary care practice activities, supported by a project team member (the principal investigator). The "Second Order Action Research" element of the study involved the project team collecting data about learning processes within participating practices. Methods included:

- Field notes of development activities (ongoing throughout the project),
- Semi-structured interviews (15 interviews),
- Meta-analysis of practice questionnaires, and
- Collection of narratives through construction of digital stories. (4)

Each of the participating practices in Phase 2 constructed a learning plan and identified a project development effort. The project team facilitated this process through:

- Semi-structured interviews with lead clinicians, practice managers, and a sample of others supporting the delivery of primary care, such as receptionists.
- Focus groups with members of the primary care team.
- Bespoke facilitation of learning within practices. Typically, this focused upon identifying and addressing challenges to engaging in collaborative and reflective practice.
- Individual coaching.

Self-Assessment. An important element in supporting the development of a learning organization is enabling review and assessment/evaluation of activity. In order to provide a tool for practice of self-assessment that was aligned with the capacity building philosophy of the MPLO project, participating practices were provided with adaptable *self-assessment questionnaire templates*. (The templates were based upon the validated survey of Garvin et al. 2008.)

Practices constructed a portfolio of their learning. Participating practices met during the course of Phase 2 to share experiences and learning within their own practice organizations. However, plans for practices to meet each other during the project and a conference to enable practices to present their portfolios to each other, fell victim to the challenge to engage in external activities. Three practices chose to develop and administer their own learning organization self-assessment tool. Practices adapted the content and questions within the tool and the method of data gathering and categorization, to better match their own perceived needs.

Administration of the self-assessment tool. The self-assessment tools were used twice by each practice, approximately six to seven months apart. Each practice used the results internally for their own reflection and self-assessment. The data was also returned to the project team. Data were entered onto a database and Excel spreadsheets constructed to analyze the data returned for each adapted questionnaire.

Analysis of results from the self-assessment tool. Data from each of six sets of returns were entered into specially constructed Excel spreadsheets that calculated arithmetic average responses overall, average by *Part*, average by *Section,* and average by *Question*. The self-assessment questionnaires formed part of the portfolio of learning that each of the four completing practices wrote as review of their learning as the project ended.

Adjustments. Fifteen practices expressed an interest in the project following the pilot study. Of these, six established an initial commitment to the project beyond the first 6 months. Four practices stayed to the end of the project submitting a portfolio

Table 15.3 Project Activity Adjustments

Structure of activities	Processes and adjustments
Baseline practices as LO	Interviews and focus groups
Identify learning need and create a learning plan proposal	Two proposals and five development areas
Supporting implementation: Practice visits for bespoke advising	Coaching (4), workshops (8), briefing (6)
Creation of website	Firewalls and practices develop intranets
Digital story workshops	1 out of 2 took place
Three meetings with all practice	None took place
Interviews and focus groups	Access problematic: 15 interviews and 8 focus groups. Telephone support

of their learning. Those practices that dropped out of the project cited a range of reasons including: illness of General Practitioners and practice managers, the demands of annual reporting for the national Quality and Outcomes Framework which has financial consequences, visits by local health authorities, and lack of time. These reflect some of the obstacles to engagement that practices identified during the pilot study and explain some of the difficulties in recruiting additional practices to the project.

Patterns. Patterns that were emerging from enthusiastic practices were that they could not sustain team engagement for voluntary initiatives because they were struggling to meet service expectations. It seemed that practices were working to their full capacity with little or no tolerance for any additional strain or demand.

Typically in empirical work, project plans have to be adjusted to take account of unanticipated events. The incidence of such challenges was very high in this project and required the project team to be particularly agile and adaptive in adjusting their supporting strategies to meet the needs of participating practices. (Table 15.3: Project Activity Adjustments.) Purposes of these adjustments were to enable practice engagement and support development. Facilitating small groups, coaching individuals, and providing telephone advice became the main forms of support.

15.8 Findings: Challenge and Needs in the Marshlands of Work-Based Learning

As previously noted, recruitment and retention to the project proved problematic and appeared to be indicative of the obstacles practices experienced in taking on commitments beyond the everyday demands of primary care. Against this backdrop, participating practices struggled to engage with the MPLO project. Heavy workloads and competing pressures on practices to be *visited* by other external bodies limited project team access to practices. Planned meetings were subject to last-minute rescheduling or cancelation. Activities, which involved people leaving their work and attending outside events with financial support, failed to recruit. Practices felt and behaved as if they had little or no capacity to engage in anything other than day-to-day service delivery. Primary care appeared to be a constantly pressurized environment.

Within participating practices development activity focused on:

- Addressing issues of respect and trust which were essential for the *improvement of communication* horizontally within clinical and nonclinical groups and vertically, moving up and down hierarchy. This was also important for developing a sense of safety to contribute and deal with sensitive or otherwise challenging issues.
- *Re-visiting what work-based learning might be* and helping those within practices to engage in reflective conversations about everyday working events and value learning from that.
- The project team arranged visits to practices to *explain how narratives* and the role of individual narratives through the creation of digital stories *could support learning and reflective practice.* As the project progressed, the work-based learning definitions appeared less useful.
- *Finding ways of recording learning that was not onerous.* The following strategies were used. The use of dictaphones, postings on practice intranets, use of meeting minutes, and the customization of surveys that each practice used to identify their progress as a learning organization.
- *Coaching individual practice team members.* Receptionists, practice managers, and nurses were included in coaching.

15.9 Primary Care Practice Projects

The four primary care practices to complete their projects each approached the work slightly differently and defined their own priorities. Practice 1 and Practice 2 submitted a proposal. Though both regarded themselves to be learning organizations, Practice 2 talked about the remaining challenges they felt they faced. Practices 3 and 4 could not see the point of the proposal. For them that was an academic construct and they just wanted to get on with the work. Practice 1 proposed finding out what teaching and training skills existed within the practice and would use this to help staff recognize those skills in themselves and build upon them. Practice 2 wanted to further change the culture within their practice and set themselves a series of objectives. Practice 3 developed a new intranet to improve practice communication and share learning. Practice 4 started a blog to improve communication and extend the use of the practice Internet to share documents and support informal learning. They also developed training activities for receptionists.

15.9.1 A Closer Look at Practice Activity: Practice 2 Vignette

The following vignette from a practice portfolio provides a snapshot of practice participation and reporting. The MPLO project team provided the headings.

> Practice 2: Profile and Rationale for joining the project.
>
> We have a history and interest in building a learning organization. Staff development, quality of service, patient safety and patient care are the four fundamental principles of our daily engagement in Primary Health Care.
>
> We became a Vocational Training Practice for clinical learners as well as being one of the first Practice Nurse (1990) and Manager (2005) training practices in the country.
>
> We have also made sustained investment in staff training and development, e.g., non-General Practitioner Associate Trainer, Continuing Professional Development (CPD) and lifelong learning to improve capacity, and to be ready to respond to and, even compete in, the changing social and economic environment.
>
> … We were inspired to join the Multi-Professional Learning Organization Project to support the development of the culture of whole team learning and, hopefully, improve the quality of care offered to patients.
>
> In the (Developing multi-professional learning organization) project, these are our goals.
>
> - Create a culture in the practice that celebrates and encourages success and innovation, a culture that recognizes and has scope for acknowledging and learning from past mistakes.
> - Protect time and space for multi-disciplinary learning in teams and as a whole practice.
> - Improve effective communication for the team and for patients at a time of National Health Service reforms.
> - Provide a safer, quality environment for our patients and for team members.
> - Be sustainable for the future.
> - Learn from others and share experiences through networking with other organizations.
>
> Summary of activities.
>
> - We extended Team Talk—weekly facilitated protected learning for reception, administration, and secretarial staff.
> - We created opportunities for team building and team working—building trust, respect, and value across professional teams, across the whole practice team and with our patients via the Patient Participation Group.
> - We organized our first extended whole practice-learning event.
> - We undertook a series of three Learning Organization Evaluation questionnaires.

Practice 2 *serves 18, 000 patients in an area of high unemployment, poverty, social deprivation, childhood obesity, teenage pregnancy, safeguarding issues, single parent families, an ageing population, chronic illness, and persons who care for the elderly. All this contributes to higher than average consultation rates and other demands on services offered by the Team.* (Extracts from Practice 2 Portfolio.)

The portfolio documents the implementation and refining of communication activities, some planned at the outset of the practice development plan and others that emerged. Two electronic forms of communication were added to the practice intranet. A daily bulletin was produced for all staff with reminders of cover arrangements for holidays, other changes to working patterns, visits to the practice from external people or agencies, and progress on actions agreed at practice meetings. The Tree of Knowledge was an interactive communication bulletin board where all staff could post social- and work-related information. This stimulated cross-practice communication and provided a means for continuing conversations about progress of the development plan and other work issues that were arising.

Team Talk became an established weekly communication and development meeting for nonclinical staff. However, following feedback from this group, a general practitioner or partner representative attended each meeting to ensure the visibility of two-way communication. Secretaries' lunch-breaks were adjusted to ensure that their attendance at these meetings were remunerated.

The practice did hold two whole practice-learning events. One was a "Significant Event Audit" workshop with an external speaker, Prof. Mike Pringle. Active in primary care development at a senior level, Professor Pringle is a past Chairman of the Royal College of General Practitioners. He complimented the practice on their open and inclusive cross-professional group conversations.

The practice was accepted onto the testing of a curriculum that formed part of the national *Productive General Practice Programme* and their self-assessment questionnaire was used as a basis for development work within that initiative. A practice manager of Practice 2 said that they had been accepted on the basis of the work undertaken with the "Developing Multi-professional Learning Organizations" project. However, the demands of the national project proved onerous as a practice manager reflected upon in their Practice portfolio:

> We hoped the [*Productive General Practice Programme*] could continue the [(kind of]) support we (were having) to develop as a learning organization. Our experience is that (initiative) is hugely demanding on us (practice managers) and on other staff. There is little of no awareness of the realities of our daily working lives and the demands this makes on our time in the context of excellence in patient care.

The portfolio concludes with an observation from a practice manager of a whole practice development meeting.

> When I was aware of laughter around the room, I looked and I could see virtually everyone was involved, smiling, nodding, laughing-and I thought- It's happening- we are getting there.

How effective was engagement with practices and what did we learn?

Three out of the completing practices focused upon improving *communication, respect for individuals and creating safe conditions in which to learn* and address sensitive or difficult issues. The fourth practice focused upon developing further as a teaching organization.

15.10 Self-assessment Insights and Legacy

Two practices claimed that they would continue to use their adapted self- assessment questionnaire after the project had finished. In the self-assessment survey categories relating to communication, trust, and working in a supportive learning environment, there tended to be a decline between the first and second administration of the survey. This dip in scores triangulated with data from interviews, focus groups, and workshops that revealed concerns, particularly among administrative staff, with their experience in these areas. As the project progressed, practices appeared to become more openly self-critical and comfortable with expressing concerns and problems.

Portfolios revealed attention spent on development processes to improve communication and a climate in which issues could be identified and addressed. For example, the development of practice intranets, away days, and weekly meetings called "Team Talk," and e-newsletters.

15.10.1 Rethinking Work-Based Learning

Initially, some of the participating practices thought the project would be delivering formal inputs about how to become a learning organization, with clear tasks and targets (as in the *RCGP Module 5: Learning Organizations, for the Quality Practice Award*). Instead, the project engaged them in thinking about what a learning organization was or might be what their own learning needs were and how they might start to address those.

However, the project team had to rethink their assumptions that facilitation would involve a neat process of:

- Moving from an understanding of learning organizations to,
- Creating a baseline of learning needs,
- Submitting a proposal for a development project related to learning needs, and
- Implementing and reviewing that project within an action research cycle.

The learning journey for practices was much more messy. The working life of the practice shaped both how their learning could be prioritized and progressed.

Practices needed to recognize and value learning from everyday work. This involved engaging in and valuing new methods of learning and ways of knowing. For example: Conversational learning that takes place at the coffee machine, finding quick and easy ways to record critical incidents, success, or good practice. These formed foci of facilitation and recognition that this kind of learning was of value. There were other development needs that addressed not just valuing different ways of knowing but understanding the importance of the cultural climate of learning, the values underpinning the learning process. This involved valuing and respecting people.

Safe spaces to learn and share. When working with individuals in private spaces, the conditions to explore difficult issues could be created. Physical privacy, a consulting room or an office, helped facilitate candid discussion. However, such private spaces were few and they were frequently in use. The physical constraints of spaces in which to learn presented challenges when addressing sensitive issues around respect for individuals, responding to error, managing angry or distressed patients or senior staff. There were other kinds of space constraints. These related to the culture of the working environment and how people were valued. Professional hierarchies, cultural values relating ways of knowing, how roles were understood and appreciated and expectations played a significant role in shaping practice engagement with the MPLO project.

The minimum essential preconditions to becoming a learning organization appeared to be: respect for individuals, a safe environment in which to speak, being *really* listened too, and diversity of thinking. Without these the vision of Senge, Argyris, Kelly et al. and Kline would be difficult to achieve. *Trust, respect for individuals and safe spaces to work and learn emerged as essential prerequisites for a learning organization.*

Implications for developing learning organizations. While participating primary care practices welcomed development which enabled them to identify and address their own priorities, the context in which they worked had implications for how they can helpfully be engaged in learning and how their learning can be appropriately assessed and valued.

In the busy and demanding context of primary care, it is important to help practices understand the commitment of engaging in work-based learning of this kind. A “learning contract” can help clarify that commitment and provide a shield to help prioritize activities that otherwise fall victim to an array of competing demands. Such a learning contract would include: *minimum time implications* to take part in the project, *minimum funding* for the learning contract, *required external meetings or events, key activities, and deadlines*. A signed contract formalizes an agreement but does not need to compromise the personalized and customized approach to learning described in this initiative.

In the new context of “Liberating the NHS” policy, with its emphasis upon continuing professional development and creation of national learning outcomes, it is prudent to anticipate the implications for this form of work-based learning and how it might assess learning, efficacy, and worth.

For primary care practices, learning and assessment needed a strong improvement of care focus. To achieve this, they needed learning and assessment approaches that understood the everyday realities of providing care and enabled them to develop a considered response rather than token compliance to changing service demands.

Implications for Assessment. The convergence of learning and performance management agendas, together with an extension of the stakeholders in healthcare education, appears to have had a significant impact on the context of assessment, *making it increasingly high stake.*

Some participating practices expressed concern that the MPLO project team might be part of a deanery agenda to scrutinize internal practice issues and processes. It took time for the project team to build trust and reassure practices that we were impartial. Even when this was established, practices remained hesitant about sharing their learning sometimes within and across practices. Performance management, with its financial and reputational impacts, had grown over the preceding decade. Had this contributed to the difficulty in recruiting practices? Did avoidance of initiatives like the MPLO project not only preserve limited resources and time but also guard against becoming vulnerable?

The MPLO project team had all worked in healthcare and primary care for some time. Something had changed with considerable effects. Practices seemed to be more hunkered and on the edge of keeping pace with daily service demands. Had policy created an accountability overload that would affect learning and assessment?

As practices prepared their portfolios, it was necessary to reassure them that they were engaging in formative assessments/judgments that were part of the action research cycle. This was not a local variation of the national "Quality Outcomes Framework" that linked self-disclosure of level of service to financial reimbursement. The practice portfolios represented their accounts and judgments of their progress in becoming a learning organization. They were primarily, self-assessments of learning to inform their future development as learning organizations. The MPLO project team hoped to learn from these about how to best support this form of continuing professional development in the future.

As a publically funded development initiative, would this approach be acceptable when continuing professional development would need to comply with national learning outcomes? The answer to that question is still unknown.

From an educational perspective, what form of assessment or evaluation would be useful, serve primary care practices and the public good?

Dealing with Practicalities, Complexities, and the Need for Consensus. Deanery and practice purposes were complex and not easily aligned. Both parties were genuinely committed to developing learning organizations. What they meant by that was not always clear or stable. For example: Some commissioners associated the development of learning organizations with an increase in formal teaching capacity; for others, it was about understanding the learning environment and how this could be enhanced to support informal learning.

Primary care practices, who were all training practices, wanted to be better able to survive service demands to reach their potential for a broader healthcare contribution. Theirs was an improvement agenda to enable them to do more whether focusing on quality enhancement or other healthcare activities. For them, learning, teaching, and assessment was more consciously, or perhaps more immediately, linked to service improvement.

The MPLO team mediated and negotiated shared expectations of the project throughout Phase 2, building the implied consensus in the initial brief of the project. This is not unusual, particularly in the early stages of innovation and development; however, it has implications for assessment and evidence of success.

The MPLO project could be seen as a pilot in learning how to engage whole practices in work-based learning with the particular focus of developing learning organizations. To determine the appropriate use of public funding, the focus of assessment or evaluation rests most appropriately with the programme/project to develop learning organizations rather than individual practices.

A closer look at assessment issues: expectations and outcomes. Participating practices were new to the action research process and still in the early stages of learning how to engage with it. This will be reflected in the kind of development projects they can or are likely to undertake and the speed of their progress. Interestingly, practices focused upon getting the basics of learning organizations in place, not from a strong theoretical perspective but to address practical areas in need of improvement. Learning and team achievements were evident in portfolios in the areas of: communication, safe spaces in which to talk and be listened too and identifying, recording, and reflecting upon successes and challenges. While these achievements had the potential to improve conditions and motivation to learn within primary care practices, they would not necessarily quickly deliver an increase in teaching capacity, one of the deanery goals.

Portfolios or parts of portfolios could serve both external audiences and an MPLO project aim to share learning beyond individual practices. However, for this to happen, the *purpose* and *audiences* portfolios that are intended to serve need to be considered, made explicit, and supported. *Boundaries between the private world of primary care practices and wider professional and public sharing need to be re-established*. Erosion of shared expectations and agreements around this threaten to frustrate relationships between different parts of the NHS and jeopardize collaborative relationships. Potential external audiences include: other primary care practices, local health authorities, and national educational commissioning bodies.

A related challenge is how primary care practices document their learning and self-assessments within portfolios. Participating practices did not want to be burdened with too much writing, particularly of an academic kind. Their preference was to integrate documenting and reporting their learning with the kind of approaches they were using in their day job, such as minutes of meetings, bulletins, postings on intranets, and snippets from coffee conversations. If the portfolios are to communicate to external audiences and not be burdensome, alternative approaches to reporting may need to be considered. Primary care practices were agreeable to sharing their self-assessments with the understanding that these would be anonymized.

Culture, values, and attitudes. Learning organizations involve values and attitudes. For working relationships to be trusting, respectful, open, and honest involves "wicked" competencies. Knight claimed:

> (Wicked competences) resist definition, shift shape and are never solved. Such soft skills are highly valued in the workplace …(They are) achievements that cannot be neatly pre-specified, take time to develop, and resist measurement-based approaches to assessment. (Knight and Page 2007)

Drawing upon assessment in higher education literature, Knight outlines the key features of assessment "sensitive to" wicked competencies. These include:

- Recognition that assessments are provisional judgments, based on evidence at-hand and they need to be represented as such.
- The design of coherent work-integrated programmes that take a progressive view of learning and dovetail learning design and assessment design.
- Assessment that engage (learners) as participants. This supports lifetime learning. Recognition that feedback is crucial and comes from multiple sources (Knight and Page 2007).

Placing assessment at the programme level could be a way forward for action research-based initiatives such as the MPLO project.

The approach proposed by Knight in "Fostering and Assessing Wicked Competencies" (Knight and Page 2007) acknowledges and accommodates the personalized and contextualized development projects undertaken by participating primary care practices. These are typical of most forms of action research development. It is an assessment approach that is appropriate to a learning-driven curriculum. In the MPLO project, practice needs and priorities were driving learning.

Placing assessment at the program level aligns with the evaluative nature of action research and focuses upon whether the curriculum supporting learning is appropriate to need and context. This provides a means of handling assessments of practice progress in climates sensitized to public scrutiny. A programme-level assessment would involve determining the extent to which the "*Development of Multi-Professional Learning Organizations*" project achieved its main aim: to develop learning organizations in primary care in the Eastern region. This forms the second-order action research element of the project, which addresses questions such as: How appropriate was the support, the resources, the design of the intervention in helping practices? What was successful? What needed to be improved? What was omitted that might have helped or was needed? Crucially, it also allows for the context in which practices are working and learning to be understood and taken into account.

Programme-level assessment provides a view of learning that anticipates a trajectory of learning. For example, primary care practices were novice users of action research and most saw themselves as either becoming or novice learning organizations. *Their successful engagement depended, in part, on the MPLO team understanding and navigating the real and pressurized realities of primary care*. Adjusting planned facilitation approaches and activities was necessary and part of an organizational response that programme-level assessment takes account off.

The first-order action research involves primary care practices engaging in the action research review and development cycle. This offers a method for scrutinizing the shifting and measurement resistant "wicked competencies." *Respect, trust and communication were defined, developed and reviewed in practice contexts. Their provisional nature was located in situations and the extent to which team members practiced and experienced them.* There is the potential for the action research process to coherently align learning and formative feedback of "wicked competencies."

The "*Developing Multi-Professional Learning Organization*" project took continuing professional development into the real, messy, and pressured realities of primary care. *In the fluid contexts of practice and policy learning had become high risk; draining time and resources from service delivery and potentially revealing vulnerabilities that may have detrimental consequences for practices.*

Caught in a transition involving the reorganization of deaneries and postgraduate educational structures, this was perhaps a particularly challenging time to engage in innovation. The recruitment and retention problems of the MPLO project were troubling because practices wanted and felt they needed to engage in this initiative but found it difficult to do so.

It was necessary for the MPLO project team to rethink how best to engage primary care teams in work-based learning. Looking at the project through a learning and assessment perspective raises issues that go beyond developing strategies to facilitate and appropriately assess learning. It also raises issues about the "Scholarship of Application," the relationship between universities and the communities they serve. How far are universities prepared to support continuing professional development that meets the perceived needs of primary care practices? How far are they prepared to rethink assessment approaches that are not onerous, not locked in the custom and practices of higher education but serve the need for accountability and assessment for improvement?

Issues and Questions for Reflection

- This study evidenced that primary care practices have changed as places in which to work and learn as a consequence of policies to modernize and reform the National Health Service (NHS). Understanding the pressurized and demand-led environment, in which cross-professional primary care teams operate within, is important. These conditions need to inform the design and implementation of work-based learning and assessment. Developing new approaches will require a rethinking of what work-based learning and work-based assessment may involve
- Essential competencies to engage in work-based learning include: respect for individuals, trust (which includes safety to deal with sensitive or difficult issues), and communication
- For assessment to support learning and service improvement for cross-professional team working, it is helpful to:
  - build a consensus around assessment criteria among stakeholders,
  - include team-based self-assessments, and
  - consider a programme- or system-based assessment component that can deal with practice and policy contexts. These contexts are very fluid within the NHS
- Reconsidering how to support collaborative team working in constantly changing clinical contexts raises issues about "The Scholarship of Application," the relationship between universities and the communities

they serve. This will require a rethinking about how to conduct work-based assessments in ways that take account of service demands

- To what extent are universities willing and able to develop work-integrated learning and assessment approaches that are team based?
- As assessment becomes increasingly high risk for primary care practices what can be done by medical and health professional educationalists to create a safe space for them to learn and engage in assessment for learning and practice improvement?

References

Argyris, M., & Schön, D. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley.

Department of Health. (1998). *A review of continuing professional development in general practice*. London: Department of Health.

Department of Health. (2010). *Liberating the NHS: Developing the healthcare workforce—from design to delivery.* www.dhgov.uk/health/2012/01/forum-response. Accessed March 6, 2012.

Department of Health. (2012). *Liberating the NHS: Equity and excellence.* (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_117353. Accessed March 7, 2012 (First published 12 July 2010).

Eastern Deanery Policies and Practice. http://www.eoedeanery.nhs.uk. Accessed March 17, 2011.

Elliott, J. (1991). *Action research for educational change*. Buckingham Open University Press.

Garvin, D., Edmonson, A., & Gino, F. (2008, March). Is yours a learning organization? *Harvard Business Review.*

hscic Quality and Outcomes Framework: Http://www.qof.ic.nhs.uk. Accessed February 24, 2016.

Knight, P., & Page, A. (2007). *Fostering and assessing wicked competencies*. Milton Keynes: Open University: http://www8.open.ac.uk/opencetl/resources/pbpl-resources/knight-2007-fostering-and-assessing-wicked-competencies. Accessed February 24, 2016.

Knight, P., & Yorke, M. (2003). *Assessment, learning and employability*. Maidenhead: Society for the Research in Higher Education, Open University Press.

McKee, A., & Watts, M. (2003). Practice and professional development plans in East Anglia: A case of politics, policy and practice. In *Work based learning in primary care* (Vol. 1, pp. 33–47). Oxford: Radcliffe Medical Press.

Quality and Outcomes Framework. http://www.qof.ic.nhs.uk. Accessed March 18, 2011.

Reason, P., & Bradbury, H. (Eds.). (2001). *Handbook of action research, participative inquiry and practice*. London: Sage Publications.

Royal College of General Practitioners QPA Essential Guidance. www.rcgp.org.uk/professional_devlopmnt/team_quality/qpa_qpa_essential_guidance.aspx. Accessed March 18, 2012.

Rushmer, R., Kelly, D., Lough, M., Wilkinson, J., & Davies, H. T. (2004a). Introducing the learning practice-I. The characteristics of learning organizations in primary care. In A. Miles (Ed.). *Journal of the Evaluation of Clinical Practice,* 10(03), 375–386 (pub Wiley-Blackwell).

Rushmer, R., Kelly, D., Lough, M., Wilkinson, J. E., & Davies, H. T. (2004b). Introducing the learning practice-II. Becoming a learning practice. In A. Miles (Ed.). *Journal of the Evaluation of Clinical Practice,* 10(03), 387–398 (pub Wiley-Blackwell).

Rushmer, R., Kelly, D., Lough, M., Wilkinson, J. E., & Davies, H. T. (2004c) Introducing the learning practice-III. Leadership, empowerment, protected time and reflective practice as core contextual conditions. In A. Miles (Ed.). *Journal of the Evaluation of Clinical Practice,* 10(03), 399–405 (pub Wiley-Blackwell).

Senge, P. (1994). *The fifth discipline* (p. 4). Fieldbook, Doubleday.

Chapter 16
Faculty Development in Assessment: What the Faculty Need to Know and Do

Ara Tekian and John J. Norcini

Abstract Most faculty in professional schools do not have formal training in education despite the fact that it is an important part of their responsibility. This is particularly true for assessment. Faculty development in assessment plays a central role in establishing accountability, and motivating and creating learning. At the core of good assessment is a well-codified body of knowledge and the need for practice. The purpose of this chapter is to outline the components of a complete faculty development program in assessment, intended for faculty in various health professions settings. Included are basic principles, methods, guidelines for blueprinting and test construction, assessor training, and scoring and standard settings. The chapter consists of the following five sections with a separate workshop for each section. (A) Steps in constructing a test and the criteria for good assessment. Criteria for good assessment is discussed and for faculty development purposes, issues related to reliability, validity or coherence, equivalence, feasibility, educational affect, catalytic effect, and acceptability are elaborated. (B) Methods and their alignment with competencies. Alignment of educational objectives with instructional and assessment methods is extremely important, and it contributes to the effectiveness of a curriculum. Examples of appropriate alignment are presented at both undergraduate and postgraduate levels. (C) Blueprinting and test construction. In the professions, the quality of the individual examinations is driven mainly by the content. Consequently, it is important to systematically sample from the domain of the competencies being assessed. Likewise, the quality of test material is paramount. (D) Assessor training. Many of the methods of assessments rely on the judgments of the faculty and other observers. The quality of these assessments is strongly influenced by the accuracy and consistency of the judges. This section focuses on how to train these assessors and includes practical methods for

A. Tekian (✉)
Faculty of Education, University of Illinois at Chicago, Chicago, IL, USA
e-mail: tekian@uic.edu

J.J. Norcini
FAIMER, Philadelphia, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change
in Professional Education 13, DOI 10.1007/978-3-319-30064-1_16

enhancing, improving, calibrating, and ensuring the quality of their judgments. (E) Scoring and standard setting. In many testing situations it is important to assign numerical scores to student performances, and to make pass/fail decisions. This section describes commonly used scoring strategies and highlights several methods for making pass/fail decisions.

Takeaways

- All faculty need to have a minimal exposure to assessment concepts.
- The assessment concepts can be delivered in a series of short workshops over time.
- Workshops should focus on basic principles, methods, guidelines for blueprinting and test construction, assessor training, and scoring and standard setting.
- It is essential that the workshops be interactive and include significant time for hands-on activities.
- The templates provided in this chapter can be modified according to the needs of the user.

16.1 Steps in Constructing a Test and the Criteria for Good Assessment: An Introduction to Assessment

Faculty is often called on to develop tests for purposes of providing feedback to students (formative assessment) or for making decisions about them (summative assessment). Without any guidance, this is a formidable task and individual faculty members often flounder or fall back on the way they were assessed. Consequently, it is often helpful to provide faculty new to the subject with a general overview consisting of the steps in constructing a test and the criteria for good assessment.

Steps in constructing a test. In constructing a test there are several steps along the way that, when taken in order, help ensure that the final product is fit for purpose. Those steps are (1) deciding on a purpose, (2) deciding on testing time and method of administration, (3) deciding on test standardization, (4) deciding on test content, (5) deciding on the item format, (6) deciding on the number of items/cases, (7) developing the test material, (8) evaluating and selecting the items, and (9) setting the passing mark (if needed). A brief introduction to these steps will raise many of the issues of importance relative to assessment.

Criteria for good assessment. Over the past 20 years, there has been a shift in what are considered the criteria for good assessment. Historically, there was a focus on reliability and validity alone. These were expanded to include educational effect, acceptability, and feasibility in the mid-1990s (van der Vleuten 1996). More recently and in recognition of the growing role of formative assessment, a consensus process expanded validity to include coherence, expanded reliability to include reproducibility and consistency, and added criteria for equivalence and the catalytic effect (Norcini et al. 2011). These criteria outline the fundamental concepts in assessment and a working knowledge of them will form the basis for an introductory faculty development workshop which lays the groundwork for future learning.

These basic concepts do not apply equally well to all testing situations. For example, they will have different importance depending on the purpose of assessment. An examination designed to make graduation decisions about a trainee will not, at the same time, produce good feedback aimed at providing and identifying particular strengths and weaknesses, guiding further study, and creating learning. Therefore, faculty development around the criteria is also an opportunity to introduce the concepts of formative and summative assessment.

Likewise, the criteria are not of equal value to all stakeholders given the same assessment. For example, the validity of the scores from a graduation examination may be of more importance to patients than how much it costs the students who take it or the institutions that administer it. The importance of the criteria will vary with the values of the stakeholders, so faculty development in this area provides an opportunity to introduce the concepts of stakeholders in assessment and their needs.

In addition to the learning about the basic concepts of assessment in the context of individual tests, faculty needs to become acquainted with the tenets of systems of assessment. The US National Research Council identifies the characteristics of such a system as comprehensive (formative and summative), coherent (aligned around the same curriculum, standards, goals), continuous, integrated into the educational system, and sufficiently resourced. An introduction to these concepts fits nicely around the criteria for good assessment.

16.1.1 Workshop 1

Title	Criteria for Good Assessment
Participants	Educators/basic or clinical sciences faculty/instructors in the health professions education
Numbers	20–36 individuals

Objectives/content	Activity	Duration (min)	Resources
Introductions	Participants introduce themselves	10	
Review of the workshop plan	Ppt presentation of objectives and list of activities	5	Handouts
Brainstorming	Present a scenario in which an assessment is needed and conduct a group discussion on the steps involved in creating it	20	Scenarios

(continued)

(continued)

Objectives/content	Activity	Duration (min)	Resources
Steps in constructing an assessment • Purpose • Testing time and method of administration • Standardization • Content	Interactive presentation	25	
BREAK		15	
• Formats • Number of items/cases • Write the items • Evaluate and select the items • Set the passing mark	Interactive presentation	20	
Identify the steps needed to construct an assessment for a particular purpose	Participants working in groups of 5–6 decide on the steps and present their work in the large group	25	2–3 scenarios
BREAK		10	
Criteria for good assessment	Interactive presentation	10	
Create a system of assessment for a particular competence	Participants working in groups of 5–6 decide on the system and present their work in the large group	25	
Take home message	Reflections and implications for home institutions	5	Scenario
Workshop evaluation	Complete feedback forms	5	Feedback forms

Duration 2 h 40 min (160 min)

Setting One large classroom, with 4–6 round tables, 5–6 individuals per table

Facilitators One or two assessment experts

Objectives By the conclusion of this workshop, participants will be able to:

1. Identify and apply the steps in constructing a test
2. Understand the criteria for good assessment
3. Appreciate the issues in constructing a system of assessment

16.2 Methods and Their Alignment with Competencies

Alignment of educational objectives with instructional and assessment methods is extremely important, and it contributes to the effectiveness of a curriculum. Sometimes, when performance of students is suboptimal, diagnosis of the curriculum might indicate a misalignment of the educational objectives or the competencies with the choice of assessment methods.

Over the past decades, a number of different assessment methods have been developed. It is critical that the chosen method should be aligned with the purpose of the test, the competence to be assessed, the nature of the examinees, the resources available, and the intended educational effect.

For example, if the purpose is formative, the competence to be assessed is clinical skills, and the resources are limited, then the workplace-based method of assessment might be appropriate. If the purpose is summative, and the clinical skills need to be assessed, then OSCEs might be a better choice. The decision about the method of assessment should follow the decisions about the purpose. The choice of the method of the assessment should flow from the purpose, competence, feasibility, and the educational effect.

During faculty development in assessment, time should be devoted for identification of assessment methods, assessment tools, and their advantages and limitations. Individuals conducting workshops on assessment and their alignment with competencies might find it useful to review Chaps. 7–9 in the "Assessment in Health Professions Education" book (Downing and Yudkowsky 2009). Chapter 7 is about constructed-responses and selected-responses formats, which contain a number of examples for each format, Chap. 8 is about observational assessment, which includes Table (7.2) about assessment goals and the corresponding assessment tools, advantages and limitations, and Chap. 9 is about performance tests. To select the appropriate assessment method, individuals need to know first about the various methods and then examine the alignment of the objectives with the assessment methods.

To elaborate on the concept of alignment, two sample objectives are selected from the "Core Entrustable Professional Activities (EPA) Curriculum Development Guide" published by the Association of the American Medical Colleges (AAMC 2014). The first is from EPA1, and the second from EPA 5 (See Table 18.1). The verb in the first objective is "demonstrate" and the therefore, the alignment with the assessment method, either using a mini-CEX or practicing with a standardized patient in an Objective Structured Clinical Examination (OSCE) station could be appropriate. The verb in the second objective is "document" and therefore, the assessment methods are different, it requires both creating a document as well as revision of medical charts and records. The purposes (formative and/or summative), the available resources, and the intended educational effect are all aligned with the assessment methods (Table 16.1).

It would be useful for faculty developers to review Bloom's taxonomy, highlighting verbs that are at the low level, such as "define," "describe," versus verbs that are at a higher level, such as "demonstrate," "create," "design," or "evaluate."

Table 16.2 illustrates the Six Accreditation Council for Graduate Medical Education (ACGME) competencies aligned with appropriate assessment methods ("1—Most preferred method"; "2—Second preferred method"). Note that one assessment method could assess several competencies, such as a knowledge tests that could assess five of the six ACGME competencies, or one competency might

Table 16.1 Examples of assessment methods, purpose, competencies, nature of examinee, resources, and educational effect

Assessment method	Purpose	Objective/competency	Nature of examinee	Resources	Educational effect
Observed clinical encounters -mini-CEX -OSCE	Formative/summative	**Demonstrate** patient-centered exam techniques that reflect respect for patient privacy, comfort, and safety (From *EPA 1: gather a history and perform physical exam*)	Final year medical student or entering residency	Standardized patients, real patients	Motivation to pass graduating competencies
Chart review/audit	Formative	Accurately **document** the reasoning supporting the decision making in the clinical encounter for any reader (From *EPA 5: document a clinical encounter in the patient record*)	Final year medical student or Entering residency	Medical records/charts	Self-assessment to accurately document patient records

Table 16.2 ACGME competencies and assessment methods

Methods (tools)/ ACGME competencies	Observed clinical encounters (mini-CEX)	Observed procedures (DOPs)	Multi-source feedback (MSF)	Knowledge tests (in-training exam)	Chart audit (CSR)
PC	2	2	2	1	2
MK	1	1	1	1	1
PBLI				2	1
SBP				2	1
ICS	1	1	1		2
Prof	2	2	1	2	

Note PC patient care; *MK* medical knowledge; *PBLI* practice-based learning and improvement; *SBP* system-based practice; ICS interprofessional communication skills; *Prof* professionalism; *DOPs* direct observation of procedural skills; *CSR* chart stimulated recall; *1* most preferred method; *2* second preferred method

be assessed by a number of assessment methods, such as patient care. The risk is when the assessment method is not aligned with the competency, such as assessing interpersonal communication skills by knowledge tests.

Workshop 2 provides a template on how to plan and organize a workshop for aligning assessment methods with competencies. The workshop includes short presentations and individual and group activities, and should be conducted in a very interactive way.

16.2.1 Workshop 2

Title Aligning assessment methods with educational objectives/ competencies

Participants Educators/basic or clinical sciences faculty/instructors in the health professions education

Numbers 20–36 individuals

Duration 2.5 hours (150 min)

Setting One large classroom, with 4–6 round tables, 5–6 individuals per table

Facilitators One or two assessment experts

Objectives By the conclusion of this workshop, participants will be able to:

1. Identify the following five components when selecting an assessment method: purpose, competence to be assessed, nature of the examinees, resources, and the intended educational effect.
2. Align objectives/competencies with appropriate assessment methods
3. Review and critique teaching units by examining the alignment of the objectives with the assessment methods

Objectives/content	Activity	Duration (min)	Resources
Introductions	Participants introduce themselves	10	
Review of the workshop plan	Ppt presentation of objectives and list of activities	5	Handouts
Brainstorming about the participants experience in alignment of objectives and assessment methods	Participants tell stories of alignment or misalignment at their institutions	15	
Short overview matching objectives with assessment methods – Sample tables with various objectives and assessment methods – Bloom's taxonomy – Discussion about importance of verbs – Five components of objective 1	Interactive presentation with ample probing of the participants about their experiences in aligning their courses and considering the five components mentioned in objective 1	20	
Group activity: Participants write two objectives (cognitive and psychomotor), then identify for each the appropriate assessment methods and the five components	Participants working in groups of 5–6: Coach various groups to measure the knowledge in the first objective and performance in the second while paying attention for the selection of the verbs in each objective	25	Flip charts
BREAK		15 (5 min per group)	Coffee/tea
Compare and contrast similarities and differences in the process of alignment; elaborate further if the alignment was not convincing	One or two members from each group present a summary of their exercise	25	
Presentation of examples from undergraduate and postgraduate levels	Participants ask questions and reflect on the appropriateness of the objectives with the methods	20	
Discussion of consequences in the absence of alignment	Participants reflect on alignment issues at undergraduate and postgraduate levels Reflect on their experiences about misalignment at their institutions and the subsequent consequences	10	

(continued)

(continued)

Objectives/content	Activity	Duration (min)	Resources
Take home message	Interactive discussion about implementation issues		
Workshop evaluation	Complete feedback forms	5	Feedback forms

16.3 Blueprinting and Item/Case Writing

In the health professions, the quality of individual examinations is driven mainly by the quality of the content. Consequently, it is important for faculty to identify the competencies to be assessed, sample test materials from them systematically, and then produce good test material.

Blueprinting. Appropriate sampling of content is typically accomplished by constructing a blueprint or table of specifications. Among the considerations in developing a blueprint are the dimensions of interest (e.g., organ systems), the source of the content (curricular objectives, graduation requirements, clinical practice), score interpretation (norm-referenced versus domain-referenced), and how much of the examination will be devoted to each area assessed.

Table 16.3 is an example of how to create a blueprint for an "Introduction to Clinical Medicine" clerkship that covers six topics. This assessment for this clerkship includes both written/knowledge and performance components. To assess the knowledge, multiple choice questions could be used, and for the performance part, OSCE stations could be designed.

If the exam consists of 100 MCQs, then the raw numbers could be evenly distributed according to the six topics. However, if the numbers were percentages, then they could be converted into numbers. For example, if the exam needs to have 50 MCQs, then the percentages could be divided by 2, and the decimals could be rounded. This blueprint could also provide you some information about the

Table 16.3 Blueprint of an introduction to medicine clerkship (content vs. clinical manifestations)

Topics\ domain	Etiology	Pathogenesis	Clinical features	Diagnosis	Management	Total
Hypertension	2	2	6	4	6	20
Diabetes	2	2	6	4	6	20
Anemia	2	2	1	5	5	15
Heart failure	2	1	2	5	5	15
Fever	3	1	3	4	4	15
Chest pain	2	2	3	4	4	15
Total	13	10	21	26	30	100

Values in each cell can refer to either the number or percentage of MCQ items

Table 16.4 Blueprint for OSCE examination (topics vs. clinical skills)

Topics\clinical skills	History	Physical	Counseling	Procedure	Total
Hypertension	1	1	1		3
Diabetes	1		1	1	3
Anemia	1	1			2
Heart failure	1	1	1	1	4
Fever	1	1			2
Chest pain	1	1	1	1	4
Total	6	5	4	3	18

Values refer to the number of OSCE stations

percentage of questions for each theme, such as 20 % for each hypertension and Diabetes, and 15 % for each of the remaining topics. Likewise, while assessing the domains, 30 % of the questions are allocated for management, 26 and 21 % for Diagnosis and management, respectively, and the rest for etiology and pathogenesis.

Concerning the assessment of clinical skills, to sample the domain and identify the clinical skills that you want to assess based on your objectives, you could prepare another blueprint or a table of specification and identify the number of stations needed. Here, the topics remain the same; however, the other dimension could become skills, such as history taking, physical examination, counseling and procedures (see Table 16.4).

After the blueprint is created, you could see that you need 18 OSCE stations. These stations could be as short as 5 min or as long as 15–30 min. However, if longer stations are designed that include patient notes, then it is advisable to reduce the number of stations. This blueprint indicates that history taking and physical examination constitute almost two thirds of the stations; for feasibility, “history” and “physical examination” could be combined into a single station. In these circumstances, the resources and faculty time available for designing and implementing the OSCE could be balanced appropriately with the test content specified in the blueprint.

In summary, regardless of different assessment methods used, the principle and theory applied in developing a blueprint applies. Faculty development sessions should emphasize such principles so that practitioners can have the flexibility to apply the test specification technique to different assessment methods.

Item/case writing. While it is essential to have a good blueprint which guides the development and selection of test material, it is equally important to write good MCQs, essays, or OSCE cases. The ability to generate good test material can be learned but contrary to general practice, it does not necessarily come with content expertise. Consequently, item- or case-writing workshops are essential for faculty and collaboration and critique should be included. Faculty would also serve to gain from an understanding of item/case statistics and how they can be used to improve test quality.

16.3.1 Workshop 3

Title	Blueprinting and Item/case writing
Participants	Educators/basic or clinical sciences faculty/instructors in the health professions education
Numbers	20–36 individuals
Duration	4 h
Setting	One large classroom, with 4–6 round tables, 5–6 individuals per table
Facilitators	One or two assessment experts
Objectives	By the conclusion of this workshop, participants will be able to:

1. Identify at least two dimensions of interest
2. Construct a blueprint for assessing knowledge and/or clinical skills
3. Create high-quality test material

Objectives/content	Activity	Duration (min)	Resources
Introductions	Participants introduce themselves	10	
Review of the workshop plan	Ppt presentation of objectives and list of activities	5	Handouts
Brainstorming about the participants' experience in blueprinting	Participants describe the challenges that they are facing in constructing blueprints	8	
Short overview of various methods and dimensions of blueprinting	Interactive presentation and discussion about blueprinting and sampling a domain for assessment	12	
Group activity: Participants create two blueprints one for knowledge tests and one for clinical examinations like OSCE	Participants working in groups of 5–6 facilitator monitors the group work to make sure that the dimensions on the blueprint and clearly specified	25	Flip charts
Characteristics and use of blueprints			
Pitfalls	Representatives from each group present their two blueprints, and the rest of the participants reflect on the presentation	20	
BREAK		15	Coffee/tea
Consequences of absence of blueprinting and lack of	Discussion	5	

(continued)

(continued)

Objectives/content	Activity	Duration (min)	Resources
producing high-quality test material			
Item/case writing: How to write good cases, MCQs, essays…	Interactive presentation and discussion	20	
Group activity: Participants develop some number of [cases, MCQs, essays…]	Participants working in groups of 5–6 develop test material, present it, and the material is critiqued by the whole group discussion	90	
Take home message	Interactive discussion about implementation issues	10	
Workshop evaluation	Complete feedback forms	5	Feedback forms

16.4 Assessor Training

Many of the methods of assessments used in the health professions rely on the judgments of the faculty and other observers. The quality of these assessments can be strongly influenced by the accuracy and consistency of the judges. Training assessors calibrates, enhances, improves, and ensures the validity and reliability of their judgments.

This training applies to all types of assessment including written or performance examinations. In the case of written assessments, such as short essay questions, a brief meeting and a sample response for each question can minimize variability in grading the essays. Similarly, in performance examinations, such as orals or observations of performance, assessor training reduces the range or spread of scores thus increasing the accuracy of the assessment. For example, in an oral examination, the assessors are trained according to a set of acceptable guidelines to standardize the difficulty and the nature of their questioning.

In performance assessment, assessors should understand the definition of the anchors in a rating scale or rubric with developmental levels and practice using them. This practice should occur before the actual examination or assessment and include observation of a series of simulated or videotaped performances each followed by sharing of judgments and a discussion of consistencies and conflicts. These careful procedures are a step toward harmonizing judgments, reducing examiner bias, increasing consistency, and enhancing the validity of the outcome.

Observing, encoding, retrieving, and evaluating performance can be challenging tasks that are prone to rating errors and biases, especially for complex skills or assessment settings. Rater training aims to improve rater performance by developing the necessary knowledge, skills, and attitudes to accurately evaluate demonstrated skills and competencies. Both novice and expert raters have been found to produce accurate and reliable ratings after rater training (Feldman et al. 2012).

Clinical competence of faculty is a crucial component of effective assessment, yet this issue has received little attention to date. Competency-based Medical Education (CBME)-focused faculty develop will need to incorporate clinical skills training with training in assessment to address important deficiencies in clinical skills. Faculty development will also need to incorporate training in the "new" competencies crucial to twenty-first century practice (Holmboe et al. 2011). Assessor training techniques such as performance dimension training and frame-of-reference training provide useful frameworks for conducing assessor training (for details, see Holmboe and Hawkins 2008).

Khera et al. (2005) in a study training examiners in pediatrics identified the following desirable attributes of the examiner:

1. Ability to use defined techniques to elicit the best performance from candidates
2. Understanding of educational theory and practice in relation to assessment
3. Have a understanding of reliability and validity
4. Be willing to accept training and regular monitoring of performance
5. Be active clinically

In clinical settings, such as the OSCE, Wilkson et al. (2003) argue that experience of the examiner is important but this experience should also be applied to station construction and not just rating students.

Tekian and Yudkowsky had summarized the essential steps involved in training examiner for Oral examinations. However, these steps apply to all the professions since they highlight the faculty development aspect for training assessors (Tekian and Yudkowsky 2009). The framework provided below could work in a variety of assessor training settings, including OSCE, patient note, mini-CEX, among others.

- **Select** examiners who are knowledgeable in the domain to be tested, and have good communication skills.
- **Orient** examiners to the exam purpose, procedure, and consequences (stakes).
- **Explain** the competencies to be assessed, types of questions to be asked and how to use any trigger material. Have examiners practice asking higher order questions.
- **Review** and **rehearse** rating and documentation procedures.
- Provide **frame-of-reference** training to calibrate examiners to scoring of different levels of responses.
- Have new examiners **observe an experienced examiner** and/or practice via participation in a simulated oral examination.
- Observe new examiners and **provide feedback**, after which an examiner is either ***invited or rejected***.
- Continue **ongoing calibration**/fine-tuning of examiners, particularly in high-stakes examinations.

Finally, training alone does not ensure high quality of results. It is also important to select assessors with care, provide them feedback on their performance, and excuse those who are unable to perform at a high standard. Taken as a whole, this process should yield a well-calibrated group of assessors.

Below is a prototype workshop that can be adapted after identifying the characteristics, profession, and level of experience of the participants or the intended target audience. If for example, the participants are nurses, the scenarios for role playing should be selected from nursing practice.

16.4.1 Workshop 4

Title Assessor Training

Participants Educators/basic or clinical sciences faculty/instructors in the health professions education

Numbers 20–36 individuals

Duration Two and a half hours (150 min)

Setting One large classroom, with 4–6 round tables, 5–6 individuals per table

Facilitators One or two assessment experts

Objectives By the conclusion of this workshop, participants will be able to:

1. Understand the desirable attributes of an examiner
2. Define and elaborate various anchors on rating scales
3. Practice how to narrow the rage of ratings for the same encounter that everyone has observed
4. Conduct frame-of-reference training for assessors
5. Utilize the eight essential steps involved in training examiners

Objectives/content	Activity	Duration (min)	Resources
Introductions	Participants introduce themselves	10	
Review of the workshop plan	Ppt presentation of objectives and list of activities	5	Handouts
Brainstorming Are assessors/raters trained at your institution for theoretical and practical exams? Elaborate	Participants reflect about their experiences concerning assessor training at their institutions. If no such training occurs, they reflect about various kinds of errors might occur	20	
Short presentation Review of literature where ratings have had a large range and had consequences in high-stakes examinations	Interactive discussion about the various findings from the literature review	15	

(continued)

(continued)

Objectives/content	Activity	Duration (min)	Resources
Presentation of various checklists and rating scales and discussion about anchors	Review of anchors and elaborate definitions of their meaning (descriptive and numerical)		
Video presentation Two to three videos where an encounter happens focused on some behavior or skills. The level of competence of the examinees should vary. Discussion about assessor calibration	Participants watch an encounter, then individually rate the performance of the examinee based on a rating scale provided to them. The ratings are then discussed in small groups of four to five individuals. Discrepancies among the group members are discussed with the larger group. This is repeated for the other videos as well	25	
BREAK		15	Coffee/tea
Brief brainstorming about desirable attributes of an examiner	All participants list desirable attributes of an examiner	10	
Discussion about frame-of-reference training	In small groups, participants provide examples of how frame-of-reference training could be organized for a faculty development session		
Group exercise: A topic that all participants could relate to, such as professionalism, is discussed in the context of frame-of-reference training	Participants in small groups complete an exercise about frame-of-reference training	25	
Short presentation about eight essential steps involved in training examiners (from Tekian and Yudkowsky). Each step is elaborated	Interactive discussion about the eight steps for training examiners	10	
Summary of the main points of the workshop and take home message	Reflections and implications for home institutions	10	
Workshop evaluation	Complete feedback forms	5	Feedback forms

16.5 Scoring and Standard Setting

In virtually all academic institutions, there is a need to summarize or score a student's performance on assessments, provide appropriate feedback, and occasionally select the pass-fail point. This section will address these three issues from the perspective of what the faculty needs to know.

a. *Summarizing performance*. In institutions committed to performance assessment, faculty needs to be trained in how to assign scores or use rubrics in a transparent and meaningful fashion. Further, it is critical that scoring be aligned with the test purpose. These summaries can be either quantitative or qualitative.

When numbers are being assigned to performances, there are two types of scores: raw scores and scaled scores. A *raw* score can be based on something as simple as the number right or as complex as a pattern of responses. A *scaled* score is the result of the application of a transformation to the raw scores to make them meaningful for a particular purpose. For example, scaled scores might express the performance of a student against all other students in his/her class or they might be used to ensure that the scores from different versions of the same test have the same meaning (i.e., equating).

Faculty should be exposed to the development of raw scores and they should also have a very general understanding of the purposes for calculating scaled scores. Beyond these fundamentals, there is not a need for broad-based faculty development in this area. However, it is important that a small number of faculty members or staff have a deeper acquaintance with these topics so they can serve as a resource for their institution.

Of course, numbers are not necessary to summarize a performance. Qualitative descriptions can be more or less effective depending on the purpose of the assessment. If this type of scoring or summarization is used, faculty development is imperative. While most of us have been exposed to the simple quantitative methods of scoring, the ability to capture and summarize a performance in words requires both an agreed set of standards and training.

b. Providing feedback. Giving feedback to students is one of the biggest drivers of learning. In a synthesis of meta-analyses for elementary and secondary education, Hattie and Timperley (2007) reported that of over 100 factors, feedback had one of the largest influences on achievement. Similar results have been found in medical education (Veloski et al. 2006; Jamtvedt et al. 2006).

Scores are one relatively limited form of feedback. More broadly, feedback is information given to trainees about their performance with the intention of guiding their learning and future performance (Ende 1983). Unfortunately, trainees often perceive that they do not receive enough feedback and faculty members are often too busy to spend time observing and providing guidance. For instance, a recent survey of general surgery residents about their experiences in the operating theater indicates that they are given preoperative goals less than 20 % of the time and they are given feedback afterwards on 37 % of occasions (Snyder et al. 2012). This is consistent with previous work in medicine indicating that residents are seldom observed (Day et al. 1990).

Research such as this has heightened interest in increasing the frequency and quality of feedback. Moreover, the recent growth in workplace-based assessment has offered an opportunity to provide it in a way that integrates knowledge, skills, and attitudes in the setting of patient care. While feasibility issues are beyond the reach of faculty development, training in how to provide feedback should be a central element of the curriculum. There are several models (e.g., the sandwich model, Pendleton's rules, 1984) for providing it that fit nicely into a workshop.

> *c. Setting standards*. In academic settings it is often necessary to set standards or make pass-fail decisions on a test. In many countries, this is done at the institutional level (e.g., 60 % correct is the pass-fail point) and applied without consideration to the test material or the students. While this might have been offered as a way ensuring that all students meet the same standards, it has an insidious and deleterious effect on examination validity. To achieve this outcome, faculty must select the test material with an eye towards what is likely to produce a reasonable pass rate rather than the content that is of importance. Further, it encourages faculty to modify the difficulty of the material they do choose to ensure the 'right' outcome. Ensuring that faculty are familiar with, and use, rational standard-setting methods will have a significant impact on examination quality.

Faculty should be exposed to the types of standards (i.e., absolute and relative), their relationships to scores, and some general information about the various methods. Beyond these fundamentals, there is not a need for broad-based faculty development in this area; training in specific methods that will be used will be important however. It is important that a small number of faculty members or staff have a deeper acquaintance with these topics so they can serve as a resource for their institution.

16.5.1 Workshop 5

Title	Scoring and Standard setting
Participants	Educators/basic or clinical sciences faculty/instructors in the health professions education
Numbers	20–36 individuals
Duration	3 h
Setting	One large classroom, with 4–6 round tables, 5–6 individuals per table
Facilitators	One or two assessment experts
Objectives	By the conclusion of this workshop, participants will be able to: 1. Understand the types of score interpretation and standards 2. Identify what makes a standard credible 3. Apply the major standard-setting method

Objectives/content	Activity	Duration (min)	Resources
Introductions	Participants introduce themselves	10	
Review of the workshop plan	Ppt presentation of objectives and list of activities	5	Handouts
Brainstorming	Large group discussion of current scoring and standard-setting practices and issues faced	15	
Types of score interpretation and standards	Interactive presentation	10	
Characteristics of a credible standard	Interactive presentation	15	
Standard-setting methods • Relative methods: judgments about test-takers • Absolute methods: judgments about individual test-takers	Interactive presentation	20	
BREAK		15	
Standard-setting methods • Absolute standards: judgments about items • Compromise methods • Applications to clinical exams	Interactive presentation	20	
Group exercise: application of a standard-setting method	Groups are given a portion of a test and asked to set standards using one or more of the methods. They present results and the group reflects on the exercise	50	Sample test(s)
Steps in setting a standard	Interactive presentation	10	
Take home message	Reflections and implications for home institutions	5	
Workshop evaluation	Complete feedback forms	5	Feedback forms

16.6 Conclusion

In summary, this chapter has reviewed the components of a faculty development program in assessment. The five components are basic principles, methods, guidelines, blue printing and test construction, assessor training and standard setting. Parts or combination of different sections could be used in developing these workshops. Information provided in this chapter should provide basic concepts and principles that educators and practitioners could use in designing effective faculty development sessions.

Issues/Questions for Reflection

- Do the workshops have an actual impact on practice?
- Which institution-specific best practices emerge after the workshops?
- What gaps do you find when applying this work?
- Do the workshops help you identify problems in the curriculum or teaching methods?

References

Association of American Medical Colleges. (2014). *Core entrustable professional activities curriculum development guide*.

Day, S. C., Grosso, L. G., Norcini, J. J., Blank, L. L., Swanson, D. B., & Horne, M. H. (1990). Residents' perceptions of evaluation procedures used by their training program. *Journal of General Internal Medicine, 5*, 421–426.

Downing, S. M., & Yudkowsky, R. (Eds.). (2009). *Assessment in health professions education*. New York and London: Routledge.

Ende, J. (1983). Feedback in clinical medical education. *The Journal of the American Medical Association, 250*, 777–781.

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high stakes simulation-based assessments. *The Journal of Continuing Education in the Health Professions, 32*(4), 279–286. doi:10.1002/chp.21156

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Holmboe, E. S., & Hawkins, R. E. (Eds.). (2008). *Practical guide to the evaluation of clinical competence*. Mosby-Elsevier.

Holmboe, E. S., Ward, D. S., Reznick, R. K., Katsufrakis, P. J., Leslie, K. M., Patel, V. L., et al. (2011). Faculty development in assessment: The missing link in competency-based medical education. *Academic Medicine, 86*(4), 460–467. doi:10.1097/ACM.0b013e31820cb2a7

Jamtvedt, G., Young, J. M., Kristoffersen, D. T., O'Brien, M. A., & Oxman, A. D. (2006). Audit and feedback: Effects on professional practice and health care outcomes. *The Cochrane Database of Systematic Reviews*, (1). Art. No.: CD000259. doi:10.1002/14651858.CD000259.pub2

Khera, N., Davies, H., Davies, H., Lissauer, T., Skuse, D., Wakeford, R., et al. (2005). How should pediatric examiners should be trained? *Archives of Disease in Childhood, 90*(1), 43–47.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M., Duvivier, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 conference. *Medical Teacher, 33*, 206–214.

Pendleton, D., Schofield, T., Tate, P., & Havelock, P. (1984). *The consultation: An approach to learning and teaching* (pp. 68–71). Oxford: Oxford University Press.

Snyder, R. A., Tarpley, M. J., Tarpley, J. L., Davidson, M., Brophy, C., & Dattilo, J. B. (2012). Teaching in the operating room: Results of a national survey. *Journal of Surgical Education, 69*(5), 643–649.

Tekian, A., & Yudkowsky, R. (2009). Oral examinations (Chap. 11). In S. M. Downing & R. Yudkowsky (Eds.), *Assessment methods in health professions education*. New York and London: Routledge.

van der Vleuten, C. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*, 41–67.

Veloski, J., Boex, J. R., Grasberger, M. J., Evans, A., & Wolfson, D. B. (2006). Systematic review of the literature on assessment, feedback and physicians' clinical performance. *Medical Teacher, 28*(2), 117–128.

Wilkson, T. J., Frampton, C. M., Thompson-Fawcett, M., & Egan, T. (2003). Objectivity in OSCE- checklists are no substitute for examiner commitment. *Academic Medicine, 78*(2), 219–223.

Chapter 17
Performance Assessment in Faculty Development Programs: Value of Action Research in Leadership Training

Ming Lee, Ronald Cohn, Jodi Cohn and Dan Osterweil

Abstract Most graduates of health professions enter practices without training in leadership. Short-term leadership training with no practical experience tends to be insufficient in preparing trainees for a leadership role. Action research consists of many components deemed important for leadership training. Assessing the process and outcomes of action research may also provide important information on leadership trainees' post-training workplace performances. Whether action research enhances leadership training and if it can be used as an assessment tool in the context of continuing medical education, however, has not been well investigated. This study examined the value of action research as extended educational mechanism and a post-training assessment tool in leadership training. A mixed method approach was used to assess the reports submitted by participants of a leadership training program documenting action research as a post-training exercise. An instrument using a 5-point Likert scale was developed to assess feasibility, implementation, and outcomes of the action research and leadership exhibited in the process. Three raters were involved in the scoring and thematic analysis of the reports to identify major themes emergent. Forty-two (46 %) of the 92 participants submitted a report. Mean scores showed that the participants demonstrated strong leadership skills in all but three areas: ability for team building, sustainability of gains for a longer term, and success in overcoming barriers to change ($M = 2.79$, 2.90, and 3.06, respectively). Four categories of themes were identified: action planning, implementation process, outcome and impact, and follow-up activity. The themes representing newly acquired knowledge and skills indicate extended

M. Lee (✉)
Faculty of Education, David Geffen School of Medicine,
University of California, Los Angeles, Los Angeles, CA, USA
e-mail: minglee@mednet.ucla.edu

R. Cohn
Ralston Consulting Group, Salt Lake City, UT, USA

J. Cohn
SCAN Health Plan, Long Beach, CA, USA

D. Osterweil
UCLA, Los Angeles, CA, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_17

educational value in action research. Those covered in training show the value of action research in sustaining the effects of training. This study thus demonstrates that action research after leadership training has value in consolidating and strengthening the training as well as assessing workplace leadership competency.

Takeaways

- Action research, in addition to functioning as an extended educational mechanism, can also serve as an assessment tool in leadership training.
- The assessment scale, developed in the study for action research, demonstrates sound reliability and validity for measuring leadership performance in the workplace.
- Rater training is a key to the successful use of the assessment scale for action research.
- Performance assessment in the context of continuing medical education needs further development.

17.1 Introduction

Leadership has long been recognized by educators as an important trait in all professionals (Cooke et al. 2010; Nohria and Khurana 2010). It has also been considered by healthcare system administrators as a vital change agent in institutional quality improvement (Cooke et al. 2010; Markuns et al. 2009). Leaders in all professions are expected to possess key leadership skills, such as teamwork, communication, consensus building, conflict resolution, and forward-looking vision, to cope with rapid and profound changes in an environment rife with financial, ethical, and profession-specific complexity (Ackerly et al. 2011; Kuo et al. 2010). Despite the argument that leaders are born and not bred, numerous educational interventions have been developed and delivered to trainees and professionals with the goal of imparting the relevant knowledge and fostering the needed skills for leadership roles (Ackerly et al. 2011; Busaro et al. 2011; Crites et al. 2008; Dannels et al. 2008; Goldstein et al. 2009; Kuo et al. 2010; Long et al. 2011; McDade et al. 2004; Wingard et al. 2004).

In healthcare professions, although some educational institutions have, in recent years, begun to implement leadership training components into undergraduate and graduate curricula (Ackerly et al. 2011; Busaro et al. 2011; Crites et al. 2008; Goldstein et al. 2009; Kuo et al. 2010; Long et al. 2011), the majority of leadership training programs has traditionally targeted mid-career health professionals through continuing education programs (Dannels et al. 2008; McDade et al. 2004; Wingard et al. 2004). However, short-term leadership training with no practical experience tends to be insufficient in transforming professionals into leaders (Markuns et al. 2009). Adult learning principles (Knowles 1984) suggest that leadership trainees may not readily apply the lessons learned from the training to their workplace until they assume a leadership role. Souba (2011) promoted an ontological model of leadership training which emphasizes the teaching of leadership "as a first-person, as-lived (or 'lived through') experience as opposed to a set of third-person theories or concepts" (p. 1242). These principles and models for leadership training suggest that "learning by doing" is a crucial element in any attempt to successfully transform individuals to leaders.

Action research, a systematic inquiry process that engages the participants in a series of "learning by doing" activities to achieve the goal of organizational changes and quality improvements, was widely popular in the business world during the 1980s and has since been vigorously adopted in the educational environment (French and Bell 1995; Mills 2000; Tomal 2010; Wamba 2006). Action research consists of many components deemed important for leadership training, such as identifying a quality improvement problem, developing an action plan for problem solving, and working collaboratively with relevant stakeholders to implement the plan and ensure successful completion of the efforts (Koshy et al. 2011; O'Brien 2001; Tomal 2010; Wamba 2006; Waterman et al. 2001). Short-term leadership training programs are likely to benefit from the incorporation of action research as part of the training activities. The quality and outcomes of post-training action research may be assessed to determine trainees' leadership competency in the workplace. As assessment in the context of postgraduate training is not well developed (Norcini 2014) and even less so in faculty development programs, action research may serve as an assessment tool for measuring leadership competency.

Healthcare organizations and medical institutions responded slowly to the trend of using action research for quality improvement. Some recent examples include Johnson et al. (2011) efforts in using collaborative interactive action research in workplace redesign, Lamus et al. (2011) adoption of action research as part of the curricular components to train medical students in family and community health, and Reed et al. (2012) use of an action research, called Personal Learning Plan (PLP), to motivate learning and assess outcomes in continuing medical education (CME). These three examples of published action research projects highlight some of the salient features of action research including critical reflection, active collaboration, and problem solving in real-world situations. They also demonstrate what carrying out an action research project may entail in leadership training.

The University of California at Los Angeles (UCLA) Multicampus Program in Geriatric Medicine and Gerontology (2011) created a program called Leadership and Management in Geriatrics (LMG) in 2002 to train health professionals in leadership and business management. The annually held LMG program has since incorporated action research as pre- and post-training exercises. Participants were asked to submit a preliminary plan before the training that included information on one quality improvement problem identified in the workplace, possible solutions, implementation steps, and anticipated barriers and problems. At the beginning of the training program, they presented their plan and received critiques and suggestions from the group.

Before the conclusion of the 2-day training program, they had a small group coaching session with a coach, usually one of the program faculty members, to help refine their action planning and lay out a timeline. They were then given a 2-month period to implement their action research and submit a report to summarize: (1) an action plan designed to solve the workplace problem identified; (2) steps for implementing the action plan, with critical analysis of the process; (3) implementation outcome and impact, and (4) lessons learned from the experience. They could receive an 8-hour CME credit if they completed the task and submitted a report. The program is open to healthcare professionals in any discipline and stage of their career. Fellows of the UCLA Geriatrics Fellowship Program also participated in the program as part of their fellowship training.

The purposes of this study are to conduct both quantitative and qualitative analyses of the action research reports submitted by the LMG program participants to assess their performance and experience in implementing action research, and to examine the value of action research as an extended educational mechanism and an assessment tool for post-training leadership performance in the workplace.

17.2 Methods

17.2.1 Participants

Attendees of the LMG program over a three-year period who submitted a report participated in the study.

The LMG program participants were asked to submit a written report (see above for the required structure) of no more than 10 pages two months after training to summarize their experience in implementing their action research. The action research exercise was designed to provide participants with an opportunity to apply what they had learned from training to a real life problem-solving situation, with the hope of strengthening and extending the effect of the training program.

17.2.2 Measures

We developed a scoring rubric based on a review of relevant literature on action research and its evaluation (Chapman 2010; O'Brien 2001; Tomal 2010) as well as appropriateness for the LGM program. The rating scale consists of four subscales including Proposed Action Plan, Implementation Process, Outcome and Impact, and Demonstration of Leadership, with three items for each subscale plus one item assessing the level of implementation complexity. The 13-item rubric used a 5-point Likert scale anchored from Poor (scored 1) to Excellent (scored 5), for the four subscales, and a 3-point scale anchored by Low, Median, and High, for the level of implementation complexity. The scale is included at the end of the chapter as an appendix.

In addition to using the scale to quantitatively rate the quality of the reports, we also conducted thematic analysis (Guest and MacQueen 2012a) to qualitatively analyze the content of the reports, with the goal of extracting common themes and leadership qualities emergent from the reports.

17.2.3 Procedures

Three raters (ML, RC, and JC), who are either medical educators, researchers, or both, were involved in the scoring. We used an adapted Delphi method (Linston and Turoff 2011) in the process of training the three raters to reach a consensus in the use of the rating scale and increase inter-rater reliability. Each rater rated a sample of three to five reports submitted by the participants not included in the study every time before a conference call. During the ensuing call, they compared their ratings of every item for each report and discussed the rationale behind their ratings. After two more such training sessions, they reached a consensus on how to use the scale and a comfortable confidence in conducting the task independently.

The same three raters also participated in the qualitative analysis of the reports by following the guidelines of thematic analysis to extract codes from the reports and look for emergent themes from the codes under each category of the reports. We used the structural coding approach (Guest and MacQueen 2012b) by following the preestablished report categories (action planning, implementation process, outcome and impact, and lessons learned) as the structured guide to code anything that demonstrated leadership in the tasks described under each category.

17.2.4 Data Analysis

To analyze the quantitative data, we computed Pearson correlation coefficients to examine inter-rater reliability. We used Cronbach's alpha to examine the internal consistency reliability of the scores. We calculated mean scores for the four subscales and the whole scale (excluding the implementation complexity item from the calculation), and conducted two-tailed Student t-tests to compare differences between the groups of fellows and practitioners and male and female participants. One-way analysis of variance was also performed to compare differences in the mean scores among three different levels of implementation complexity. We used the Fisher exact test to compare differences in the report submission rate between gender and career status (fellow vs professional) groups. All analyses were conducted by using IBM SPSS Statistics 20 (IBM Corp., Armonk, NY).

To conduct the qualitative analysis of the action research reports, we classified all codes and supporting quotes from the reports into several major themes which were then grouped into four categories. We discussed appropriate ways of interpreting the analysis results and their implications for training effects. This study was approved by our campus Institutional Review Board (IRB).

17.3 Results

Forty-two (46 %) of the 92 participants of the LMG program over a three-year period submitted a report. All (100 %) the reports submitted followed the required structure and described a project qualified as action research. The projects implemented covered a wide range of quality improvement issues, such as the design and implementation of an office visit summary sheet to improve healthcare efficiency in a geriatric clinic, the development and implementation of a didactic forensic psychiatry curriculum for a geriatric psychiatry fellowship, the establishment of a geriatric consultation and caregiver education service at a medical center, and the creation of interdisciplinary geriatric rounds in a hospital. All the participants who submitted a report claimed the completion of their action research projects, although some of the project activities (e.g., a year-long didactic curriculum, hospital interdisciplinary rounds) continued beyond the report submission deadline. Table 17.1 shows the demographic characteristics of the study participants and those of the LMG program participants who did not submit an action research report. The Fisher exact test showed that a significantly ($p < 0.001$) higher percentage of geriatrics fellows (76 %) submitted a report than that of healthcare professionals (29 %). There was no significant difference between gender groups in the report submission rate (39 % and 52 % for males and females, respectively).

17.3.1 Quantitative Findings

The internal consistency reliability estimated by the Cronbach's alpha for the entire scale was 0.96. The inter-rater correlation coefficients for the mean scale scores ranged from 0.74 to 0.83. The mean scores of each item and the total scale and subscales for the fellow and professional groups are shown in Table 17.2. The professionals significantly outperformed the geriatrics fellows in both the Leadership ($p < 0.01$) and Implementation ($p < 0.05$) subscales, including all three aspects of the Leadership subscale (articulation of a vision, setting priorities, and team building)

Table 17.1 Demographics of the leadership and management in geriatrics program participants who submitted a report (study participants) and those who did not

Group	Submitted a report N (%)	No report N (%)	Total participant N (%)
Gender			
M	18 (43 %)	28 (56 %)	46 (50 %)
F	24 (57 %)	22 (44 %)	46 (50 %)
Status			
Fellow	25 (60 %)	8 (16 %)	33 (36 %)
Professional	17 (40 %)	42 (84 %)	59 (64 %)

There were significantly ($p < 0.001$) more fellows than professionals who submitted a report, but no significant difference was found between the gender groups in the report submission rate

Table 17.2 Comparison of mean item and subscale scores of fellow and professional groups

Item	Total N = 42 Mean (SD[a])	Fellow N = 25 Mean (SD[a])	Professional N = 17 Mean (SD[a])	t-test
Proposed action plan				
Appropriateness	4.27 (0.68)	4.23 (0.62)	4.33 (0.77)	0.50
Usefulness	4.16 (0.68)	4.12 (0.61)	4.22 (0.80)	0.44
Timeliness	3.73 (0.75)	3.64 (0.64)	3.86 (0.89)	0.95
Subscale	4.05 (0.60)	4.00 (0.51)	4.14 (0.73)	0.74
Implementation process				
Stakeholder engagement	3.61 (1.24)	3.31 (1.24)	4.06 (1.13)	2.00
Effectiveness	3.41 (1.18)	3.08 (1.19)	3.90 (0.99)	2.34*
Success in overcoming barriers	3.06 (1.31)	2.83 (1.29)	3.41 (1.29)	1.44
Subscale	3.36 (1.16)	3.07 (1.16)	3.79 (1.05)	2.04*
Outcome and impact				
Achievement of short-term gains	3.44 (1.32)	3.19 (1.34)	3.82 (1.24)	1.56
Sustainability of gains	2.90 (1.35)	2.69 (1.16)	3.22 (1.57)	1.24
Cost-effectiveness	3.63 (1.12)	3.44 (1.25)	3.92 (0.84)	1.39
Subscale	3.33 (1.20)	3.11 (1.20)	3.65 (1.15)	1.47
Demonstration of leadership				
Articulation of a vision	3.63 (1.05)	3.31 (0.99)	4.10 (0.98)	2.56*
Setting priorities	3.55 (0.99)	3.25 (0.98)	4.00 (0.83)	2.60*
Team building	2.79 (1.25)	2.40 (1.20)	3.37 (1.12)	2.65*
Subscale	3.32 (1.01)	2.98 (0.96)	3.82 (0.88)	2.87**
Total scale	3.52 (0.92)	3.29 (0.89)	3.85 (0.90)	2.01

[a] *SD* Standard Deviation

*$p < 0.05$

**$p < 0.01$

and the aspect of effectiveness in implementing steps of the Implementation subscale. The mean scale scores for both the fellow (3.29) and professional (3.85) groups and those of the subscales for each group showed room for growth. When comparing the mean scores made by groups of participants who implemented action research with different levels of complexity, we found significant ($p < 0.05$) differences among the three groups in the total scale and all the subscales except the Proposed Action Plan subscale. These results are shown in Fig. 17.1, where the High complexity group scored significantly ($p < 0.01$) higher than the Low group on all the subscales, but only significantly ($p < 0.05$) higher than the Median group on the Leadership subscale, whereas the Median group showed a significantly ($p < 0.05$) higher mean score than the Low group only on the Leadership subscale. In addition, significantly ($p < 0.001$) more professionals than fellows (41 % vs. 0 %) implemented action research that involved interdepartmental efforts (e.g., developed and implemented interdisciplinary team care for a newly established geriatrics ward), whereas more fellows than professionals (48 % vs. 18 %) implemented action

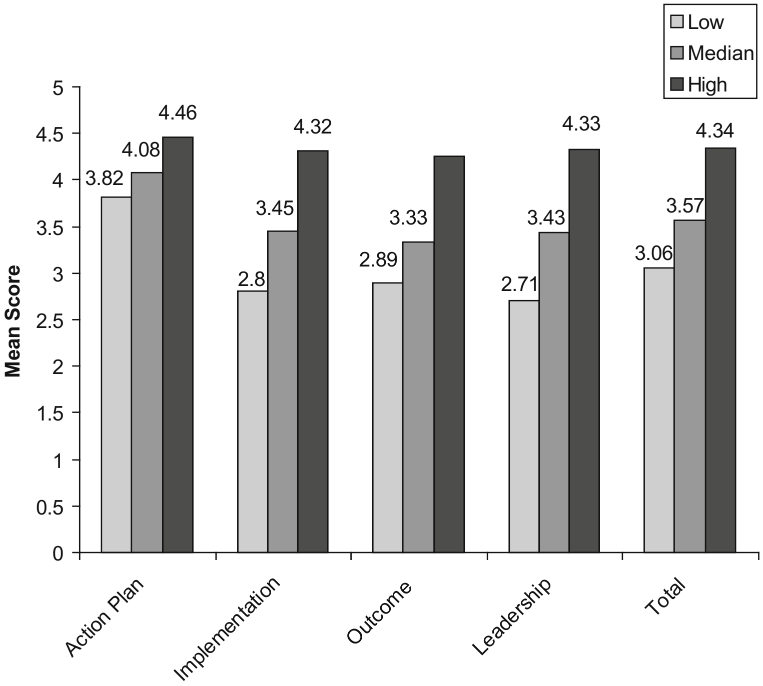


Fig. 17.1 Comparison of mean scale and subscale scores among groups implementing an action research with different levels of complexity (The mean scores of the three groups were significantly ($p < 0.05$) different from each other in the total scale and every subscale except the Proposed Action Plan subscale, with the High complexity group scoring significantly ($p < 0.01$) higher than the Low group in all areas but only significantly ($p < 0.05$) higher than the Median group on the Leadership subscale, whereas the Median group scored significantly ($p < 0.05$) higher than the Low group only on the Leadership subscale and the three levels of implementation complexity were defined as: Low—required only individual effort; Median—required team effort; and High—required cross-department or system effort)

research that required only individual effort (e.g., designed and implemented an office visit summary sheet for patients and office staff). There was no significant difference between the gender groups in the mean scale or subscale scores, nor in the complexity level of the action research implementation.

17.3.2 Qualitative Findings

We identified a number of themes from the action research reports received. The themes helped to provide insights into the leadership competency revealed in the reports. The themes were then grouped into four categories including the first three categories (Action Planning, Implementation Process, and Outcome and Impact) of the action research report and a category labeled “Follow-up Activity” to include any thoughts or plans about continuing or sustaining the action research efforts. This last category was added to replace the last category of the report, lessons learned, as the information reported in lessons learned conceptually belonged to one of the four categories, and was thus regrouped under the relevant category. The categories and themes identified are presented below, with a brief explanation and supporting quote(s) following each theme.

17.4 Action Planning

17.4.1 Theme 1. Make the Plan Feasible

Participants recognized the importance of making a small and realistic plan that could be delivered within the short timeframe that they were given. One participant also pointed out the importance of using positive language in a project proposal for easy buy-in.

> It is crucial to set small realistic goals along the way rather than trying to accomplish very broad objectives.
>
> The language used in the proposal needs to be carefully studied so that the proposal is viewed as an advantageous and non-threatening project for the establishment.

17.4.2 Theme 2. Understand the Context

Several participants learned the importance of gathering background information or conducting needs assessment to understand the context for action research.

> Rather than coming in as a new person and laying out a set plan for envisioned changes, I gathered background information to have a better understanding of the context within which I am working.
>
> I did a needs assessment… to look at clinical areas in which they wanted more assistance.

17.4.3 Theme 3. Clarify Challenges Associated with the Identified Quality Improvement (QI) Problem

Participants felt that clarifying potential challenges associated with their identified QI problem was helpful in finding solutions.

> This… challenged me to identify system problems in healthcare and thinking through ways to improve the way we deliver healthcare to our patients.
>
> Once I realized what I was really trying to achieve, it became clearer what my goals were and how I wanted to address this problem.

17.4.4 Theme 4. Seize the Right Timing

Although the implementation of action research required by the LMG program had a limited timeframe, several participants identified from their experiences proper timing as a crucial factor in any activity designed to instill organizational changes.

> I found that it can be easier to implement changes when an organization is going through restructuring as the culture is in a state of transition and more willing to make changes where needed rather than maintaining the status quo.

17.5 Implementation Process

17.5.1 Theme 1. Engage Key Stakeholders

Once the action research project was ready to be implemented, participants, in general, quickly learned that they needed to engage all key stakeholders in their projects. They needed to not only personally meet with key stakeholders, but achieve buy-in from them.

> You should know the right person to talk to if you want to get things done correctly.

> The rationale behind enlisting senior faculty members to become seminar facilitators was primarily to ensure buy-in from the department.

In some cases, it was also determined to be crucial to solicit support from the bottom-up. Some strategies identified included promoting mutual success through developing ownership among key players and publicly recognizing and appreciating contributions by them.

> I have learned that it is very important to have support from key players even before you present a proposal to the decision makers.
> We invited all nurses on both shifts to an open systems meeting for any innovative input to gain ownership of the mission.

> Appreciation and recognition can be fundamental to continuing efforts for a larger success down the road.

17.5.2 Theme 2. Develop Consensus Among Key Stakeholders

After engaging key stakeholders and achieving buy-in from them, participants usually found that articulating a vision and aligning different viewpoints among key stakeholders to reach a consensus of shared goals was an important next step. They also learned to use various strategies to help reach the goal including negotiation, standardization, development of group value and identity, and recruiting a strong leader.

> I met with PT/OT supervisors to build consensus that prior system was inadequate and to create a sense of urgency for change.
>
> Compromise and change go hand in hand. Without compromise, no effective change can be implemented.
>
> A strong leader needs to be goal-oriented to redirect the committees' focus… Without a strong leader no project can get accomplished.

17.5.3 Theme 3. Recognize Change as a Nonlinear Process

Throughout the implementation process, participants realized that solutions to their identified problems might not be reached through a linear process. Rather, they usually found other barriers in the way which needed to be dealt with first or caused them to redirect their route for problem solving.

> Often times additional problems are identified which directly influence the goals and solutions for the initial project and are necessary to consider also.
>
> I quickly realized that solutions for a single problem can often lead to other problems being identified which may also need modification.

17.5.4 Theme 4. Recognize the Value of Mentoring

The participants of the LMG program were each assigned a mentor to work with in implementing their post-training action research. Many of them expressed strong appreciation of their mentor's guidance in that process.

> I have a much richer and more specific appreciation of the value of a mentor.
>
> Having a superb mentor makes all the difference.

17.5.5 Theme 5. Cultivate Characteristics of Change Agents

A couple of the participants also pointed out the importance of developing helpful characteristics when working as a change agent. Those characteristics need to be cultivated with mindfulness and self-awareness.

> I have discovered that dealing with the external leadership and management challenges begins with me internally.

17.6 Outcome and Impact

17.6.1 Theme 1. Evaluate Outcomes

To complete the implementation of the proposed action research, some of the participants conducted an evaluation to measure the effectiveness of their projects. The majority of them, however, did not incorporate an evaluation plan in their proposal. One participant pointed out this omission as a drawback of her project.

> I learned that I made a mistake by having no method of evaluation built into my intervention.

17.6.2 Theme 2. Expect Unexpected Outcomes

Some participants also reported a number of unexpected outcomes as a result of implementing their LMG projects. They reported these as additional gains to their post-training exercise.

> In the process, not only was I able to implement my project, but was also given the opportunity to become more involved in the curriculum process for the fellowship, which was my original impetus.

Some of them were also impressed by the substantial differences they made after inducing what they thought to be simple changes to their organization.

> Simple modifications to existing processes have the potential to make a huge difference.

17.7 Follow-up Activity

17.7.1 Theme 1. Identify Next Steps and Mechanisms for Sustainability

Some participants went beyond simply reporting their current project to contemplating future steps or mechanisms to sustain project outcomes. The efforts aimed to extend the outcomes to other related areas or to potential recipients for a longer period of time.

> [The results] will be used as a springboard to analyze further admissions (and discharges) at the unit, revise the criteria for GEM admission if necessary, and implement these criteria.

> Mechanisms by which key information can be constantly updated and disseminated throughout the year were also discussed with the incoming program director.

17.8 Conclusion

We developed a rating scale and conducted a study using both quantitative and qualitative analyses to assess the process and outcome of action research designed and implemented by the participants of the Leadership and Management in Geriatrics program over a three-year period. Our goal was to examine the participants' performances and experiences in implementing action research and to evaluate the value of action research as both an extended educational mechanism in leadership training and an assessment tool for post-training leadership competency in the workplace. The rating scores on all four assessed domains and the themes emergent from the reports submitted by the study participants confirm the educational value of action research in leadership training; that is, action research is a valuable "learning by doing" mechanism which not only consolidates the lessons learned from the training but also adds new learning experiences through workplace application. Both the quantitative and qualitative findings of action research also provide reliability and validity evidence for its use as a tool for assessing leadership competency in the clinical setting.

The quantitative findings demonstrated high internal consistency reliability and sound inter-rater reliability of the rating scale. We believed that the adapted Delphi method, used in the process of developing rating consensus among the three raters, was instrumental in yielding relatively high inter-rater reliability for the scale. The findings that the participating geriatrics fellows scored significantly lower than practicing professionals in all three aspects of leadership assessed (i.e., vision articulation, priorities setting, and team building) and the effectiveness of implementing their action research supported the known-group validity of the instrument. The fellows, given their prolonged training period and limited working experience in comparison to the professionals, were expected to demonstrate less leadership proficiency than the professionals. The findings that those participants who implemented action research with high complexity received significantly higher ratings on the whole scale and across all subscales except the Proposed Action Plan subscale further demonstrated the known-group validity of the rating scale. It is conceptually reasonable to expect those who designed and implemented an action research requiring them to work collaboratively with colleagues from other departments to exhibit stronger leadership competency in their reports than those who implemented a less complex action research focusing more on individual effort. The lack of differences in the mean scores of the Proposed Action Plan subscale is likely due to the fact that all of the participants had received coaching during training to modify their original plans to make them more applicable. The rating scores collected from this scale were therefore reliable and valid.

The mean scores showed that the participants generally demonstrated strong leadership skills in all but three areas: success in overcoming barriers to change, sustainability of gains for a longer term, and ability for team building. These findings were likely a result of the fact that the participants were given only a two-month period to implement their action research and document their experiences. New ideas usually take time for absorption and acceptance, and sustainability requires all sorts of support which may not build up in a short timeframe. However, the participants as a group, and especially the professionals were able to articulate a vision with stakeholders and set priorities to achieve short-term gains in a cost-effective manner. The mean scale and subscale ratings indicated that there was room for participants to grow in leadership development. The findings, on the other hand, could also indicate that a longer period for action research implementation might be needed in order to more accurately assess leadership competency.

The qualitative analysis of the reports revealed a variety of themes related to leadership quality and skills. Cross-checking these themes with the training program agenda, we found that some of the emergent themes were taught in the program. Themes not emphasized in training but realized by the participants through their action research experiences included seizing the right timing to implement changes, recognizing that change is not a linear process, recognizing the value of mentoring, expecting unexpected outcomes, and identifying mechanisms for sustainability. The newly acquired knowledge represents a learning experience beyond the original leadership training, indicating extended educational value in action research. The themes that were covered in training were also experienced by the participants through their action research implementation. This latter outcome indicates that the learning from training was consolidated when the participants applied what they had learned to their workplace. Findings of both types of themes revealed in the reports support the prospect that the incorporation of an action research exercise after leadership training is an effective way of connecting theory and practice to enable trainees to become critical and creative in problem solving in the workplace, deepen their understanding of theory and application, and further develop leadership competency. In short, action research extends and strengthens the effects of leadership training. As long-term training is not feasible for healthcare practitioners and many professionals in the other fields, a short-term program incorporated with a post-training action research should prove to be valuable and applicable for cultivating more leaders to face the challenges of an increasingly complex healthcare and other professional systems.

Our study has several limitations. Firstly, it was conducted in one institution under one program. Other institutions or programs using different designs or in different content areas may produce different results. Secondly, less than half of the program participants submitted an action research report. Although there was no difference in gender representation among those who submitted and those who did not, there were significantly fewer professionals than fellows who submitted a report. Given their current trainee status, fellows tended to have a lower capacity than professionals to implement complex action research projects that required more than individual effort. A modest report submission rate and the higher representation of fellows in the study may affect the generalizability of the findings. Thirdly, although the three raters did not have knowledge of the participants' status (fellow versus professional) before reading their reports, it was usually not difficult to determine while reading through the reports. This lack of complete blindness may create bias in rating, but many other features of the implemented action research assessed by the study may supersede the influence of knowing the status. Lastly, our training program allowed only two months for the participants to implement their action research. Such a short timeframe is likely to limit the effect and outcomes of action research in the demonstration of leadership competency. Nonetheless, our study has already shown the value of action research in leadership training.

We continue to annually offer the LMG program. A number of measures have been undertaken to improve the collection of action research reports, including moving the program to earlier in the year to increase the time for action research implementation, and having a faculty mentor closely follow up with each participant after training. We anticipate that the changes will help increase the rates of action research implementation and report submission. By reporting the insights gleaned from this study, we hope to facilitate future investigators in the examination of the value of action research in leadership training for other levels of professional trainees, especially those in undergraduate and graduate education, as well as professionals in fields other than health care.

Questions for Reflection

- What other components may be included in the assessment of workplace leadership competency using action research, in addition to Action Planning, Implementation Process, Outcome and Impact, and Demonstration of Leadership?
- What is a reasonable timeframe for assessing post-training leadership competency in the workplace, given its diverse dimensions as shown in action research?
- How can the performance assessment scale, developed in the study, be converted to an observational measurement of leadership competency using action research?
- What other performance assessment tools may be developed for faculty development programs?

Appendix

Performance assessment scale for action research in leadership training

Performance aspect	Poor	Below average	Average	Above average	Excellent	Score
Proposed action plan						
• Appropriateness for the environmental context and problem identified	1	2	3	4	5	
• Usefulness of the intervention	1	2	3	4	5	
• Timeliness of the intervention	1	2	3	4	5	

(continued)

(continued)

Performance aspect	Poor	Below average	Average	Above average	Excellent	Score
Implementation process						
• Engagement of key stakeholders	1	2	3	4	5	
• Effectiveness in implementing steps for proposed change	1	2	3	4	5	
• Success in overcoming barriers/resistance to change	1	2	3	4	5	
Outcome/impact						
• Achievement of short-term gains/successes	1	2	3	4	5	
• Sustainability of gains for a longer term	1	2	3	4	5	
• Cost-effectiveness of the action	1	2	3	4	5	
Demonstration of leadership						
• Ability to create and articulate a vision	1	2	3	4	5	
• Ability to set priorities and directions	1	2	3	4	5	
• Ability to build and inspire a team to achieve the vision	1	2	3	4	5	
Total score						
Level of implementation complexity						
Low—Individual effort	Low		Median		High	
Median—Team effort						
High—Cross-Department/System effort						

References

Ackerly, D. C., Sangvai, D. G., Udayakumar, K., Shah, B. R., Kalman, N. S., Cho, A. H., et al. (2011). Training the next generation of physician-executives: An innovative residency pathway in management and leadership. *Academic Medicine, 86*, 575–579.

Busaro, J. O., Berkenbosch, L., & Brouns, J. W. (2011). Physicians as managers of health care delivery and the implications for postgraduate medical training: A literature review. *Teaching and Learning in Medicine, 23*, 186–196.

Chapman, A. (2010). Training programme evaluation: Training and learning evaluation, feedback forms, action plans and follow-up. Retrieved August 3, 2010, from http://www.businessballs.com/trainingprogramevaluation.htm

Cooke, M., Irby, D. M., & O'Biren, B. C. (2010). *Educating physicians: A call for reform of medical school and residency*. San Francisco, CA: Jossey-Bass.

Crites, G. E., Ebert, J. R., & Schuster, R. J. (2008). Beyond the dual degree: Development of a five-year program in leadership for medical undergraduates. *Academic Medicine, 83*, 52–58.

Dannels, S. A., Yamagata, H., McDade, S. A., Chuang, Y., Gleason, K. A., McLaughlin, J. M., et al. (2008). Evaluating a leadership program: A comparative, longitudinal study to assess the impact of the Executive Leadership in Academic Medicine (ELAM) program for women. *Academic Medicine, 83*, 488–495.

French, W., & Bell, C. (1995). *Organization development: Behavioral science interventions for organization development*. Englewood Cliffs, NJ: Prentice-Hall.

Goldstein, A. O., Calleson, D., Bearman, R., Steiner, B. D., Frasier, P. Y., & Slatt, L. (2009). Teaching advanced leadership skills in community service (ALSCS) to medical students. *Academic Medicine, 84*, 754–764.

Guest, G., & MacQueen, K. M. (Eds.). (2012a). *Applied thematic analysis*. Thousand Oaks, CA: Sage.

Guest, G., & MacQueen, K. M. (2012b). Themes and codes. In G. Guest & K. M. MacQueen (Eds.), *Applied thematic analysis* (pp. 49–78). Thousand Oaks, CA: Sage.

Johnson, P. A., Bookman, A., Bailyn, L., Harrington, M., & Orton, P. (2011). Innovation in ambulatory care: A collaborative approach to redesigning the health care workplace. *Academic Medicine, 86*, 211–216.

Knowles, M. S. (1984). *Andragogy in action: Applying modern principles of adult learning*. San Francisco, CA: Jossey-Bass.

Koshy, E., Koshy, V., & Waterman, H. (2011). *Action research in healthcare*. London, UK: Sage.

Kuo, A. K., Thyne, S. M., Chen, H. C., West, D. C., & Kamei, R. K. (2010). An innovative residency program designed to develop leaders to improve the health of children. *Academic Medicine, 85*, 1603–1608.

Lamus, F., Jaimes, C., & Serrano, N. (2011). Understanding community health medical education through action research. *Medical Education, 45*, 509–510.

Linstone, H. A., & Turoff, M. (Eds.). (2011). The Delphi method: Techniques and applications. Retrieved February 2, 2011, from http://is.njit.edu/pubs/delphibook/delphibook.pdf

Long, J. A., Lee, R. S., Federico, S., Battaglia, C., Wong, S., & Earnest, M. (2011). Developing leadership and advocacy skills in medical students through service learning. *Journal of Public Health Management and Practice, 17*, 369–372.

Markuns, J. F., Culpepper, L., & Halpin, W. J. (2009). A need for leadership in primary health care for the underserved: A call to action. *Academic Medicine, 84*, 1325–1327.

McDade, S. A., Richman, R. C., Jackson, G. B., & Morahan, P. S. (2004). Effects of participation in the Executive Leadership in Academic Medicine (ELAM) program on women faculty's perceived leadership capabilities. *Academic Medicine, 79*, 302–309.

Mills, G. (2000). *Action research: A guide for the teacher researcher*. Upper Saddle River, NJ: Prentice-Hall.

Nohria, N., & Khurana, R (Eds). (2010). *Handbook of leadership theory and practice*. In An Harvard Business School Centennial Colloquium. Boston, MA: Harvard Business School.

Norcini, J. J. (2014). Workplace assessment. In T. Swanwick (Ed.), *Understanding medical education: Evidence, theory and practice* (2nd ed., pp. 279–292). Oxford, UK: Wiley.

O'Brien, R. (2001). An overview of the methodological approach of action research. In Roberto Richardson (ed.) *Theory and practice of action research*. Retrieved September 24, 2013, from http://www.web.net/~robrien/papers/arfinal.html

Reed, V. A., Schifferdecker, K. E., & Turco, M. G. (2012). Motivating learning and assessing outcomes in continuing medical education using a personal learning plan. *Journal of Continuing Education in the Health Professions, 32*, 287–294.

Souba, W. (2011). A new model of leadership performance in health care. *Academic Medicine, 86*, 1241–1252.

Tomal, D. R. (2010). *Action research for educators* (2nd ed.). Plymouth, UK: Rowman & Littlefield Education.

University of California at Los Angeles Multicampus Program in Geriatric Medicine and Gerontology. (2011). Leadership and Management in Geriatrics (LMG) Program. Retrieved September 1, 2011, from http://geronet.ucla.edu/lmg

Wamba, N. G. (2006). Action research in school leadership program. *Academic Exchange Quarterly, 10*, 51–56.

Waterman, H., Tillen, D., Dickson, R., & de Koning, K. (2001). Action research: A systematic review and guidance for assessment. *Health Technology Assessment, 5*(23).

Wingard, D. L., Garman, K. A., & Reznik, V. (2004). Facilitating faculty success: Outcomes and cost benefit of the UCSD National Center of Leadership in Academic Medicine. *Academic Medicine, 79*, S9–S11.

Chapter 18
Assessing Competence in Medical Humanism: Development and Validation of the ICARE Scale for Assessing Humanistic Patient Care

Ming Lee, Paul F. Wimmers and Cha Chi Fung

Abstract Although the human dimensions of health care have been incorporated into medical education, how students perform in those areas remains unclear. One potential reason is the lack of reliable and valid instruments to assess humanistic performance in medical trainees. This study developed and examined a 15-item, 5-point Likert scale designed to assess medical students' performance in Integrity, Compassion, Altruism, Respect, and Empathy (ICARE). Fifty medical students' videotaped performance in an Objective Structured Clinical Examination (OSCE) station were rated by three investigators. Cronbach's alpha, intraclass correlation (ICC), and the generalizability (G) study were conducted for reliability examination. To examine validity, factor analysis was conducted to explore the latent structure of the scale, and the correlations of the scale scores with four external criterion measures were calculated. Psychometric findings provided support for internal consistency, inter-rater and reproducibility reliability as well as construct validity of the scale scores. Criterion-related validity remains to be further investigated. The mean scores for the scale (3.74) and subscales (ranging from 3.44 to 4.08) showed room for the students to grow. The mean scores on the Respect and Integrity subscales were significantly ($p < 0.05$–0.001) higher than those of the other subscales. Students' humanistic performance in a simulated clinical setting appeared to differ from their self-perceived patient-centeredness and empathy. While patient-centeredness and professionalism have received good attention in medical education, resulting in better student performances in those areas in the study, compassion and altruism remain in need of greater emphasis.

M. Lee (✉)
Faculty of Education, David Geffen School of Medicine,
University of California, Los Angeles, Los Angeles, CA, USA
e-mail: minglee@mednet.ucla.edu

P.F. Wimmers
University of California, Los Angeles, CA, USA

C.C. Fung
Keck School of Medicine of USC, Los Angeles, CA, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_18

Takeaways

- Humanism is conceptually related to, but different from, patient-centeredness and professionalism.
- The ICARE scale, a short Likert scale assessing humanistic behaviors and attitudes from the domains of Integrity, Compassion, Altruism, Respect, and Empathy, demonstrates sound reliability and construct validity.
- The moderate overall mean score and relatively lower scores on the Compassion and Altruism domains, as found in this study, shed light on potential points of future curricular emphasis with an aim o cultivating humanistic practitioners.
- A modified Delphi method helps train raters in using scoring rubrics and increase inter-rater reliability.

18.1 Introduction

The profession of medicine holds its roots in the Hippocratic value to heal the sick via a humanistic manner. Students at most medical schools in the United States, nowadays, begin their first year of training with a White Coat Ceremony, in which they cite the Hippocratic Oath and learn the responsibilities and expectations their society has imposed on those wearing a white coat (Stern and Papadakis 2006). In 1983, the American Board of Internal Medicine (ABIM) (1983) adopted four principles in emphasizing the humanistic qualities of internists. The principles specify three essential humanistic qualities of integrity, respect, and compassion, and demand candidates for ABIM certification to meet high standards of humanistic behavior in their professional lives. The Liaison Committee on Medical Education (2011) has also recognized the teaching of humanistic values as an essential component of medical education.

The word humanism carries many connotations. For example, humanism has frequently been associated with professionalism (Cohen 2007; Goldberg 2008; Stern et al. 2008; Swick 2007). It is considered as either sharing a number of components of and being fully complementary with professionalism (Cohen 2007; Swick 2007), or being a totally different construct that entails a set of beliefs which direct and lead to different sets of human conduct (Goldberg 2008; Stern et al. 2008). Rather than considering humanism and professionalism as dual values, the ABIM (1995) holds a position that states humanism is central to professionalism, enveloping the former within the latter. This view of humanism as an integral part of professionalism is shared by Gold and Gold (2006), founders of the Arnold Gold Foundation that initiates the White Coat Ceremony, who declare that humanism is the central aspect of professionalism.

Humanism has also been linked to patient-centeredness. Branch and his colleagues (Branch et al. 2001) defined humanism as "the physician's attitudes and actions that demonstrate interest in and respect for the patient and that address the patient's concerns and values" (p. 1067). The Gold Humanism Honor Society elaborated further along this line by defining humanism as "those attitudes and behaviors that emanate from a deep sensitivity and respect for others, including full acceptance of all cultural and ethnic backgrounds. Further, humanism is exemplified through compassionate, empathetic treatment of all persons while recognizing each one's needs and autonomy" (Arnold Gold Foundation 2005, p. 5). Miller and Schmidt (1999) also described a humanistic physician as one who considers the influence of patients' social, cultural, spiritual, and emotional experiences when caring for them. The conceptual framework adopted by these medical educators and organization for medical humanism shows congruence with that of patient-centeredness (Bever et al. 2010; Davis et al. 2005; Institute of Medicine 2001; Laine and Davidoff 1996; Lévesque et al. 2013; Mead and Bower 2000). In fact, the medical humanism movement that emerged in medicine over the past several decades inspired a patient-centered approach to healthcare by focusing on patients' values, goals, and preferences for medical decisions (Hartzband and Groopman 2009; Laine and Davidoff 1996; Lévesque et al. 2013).

In addition to bearing conceptual resemblance with professionalism and patient-centeredness, medical humanism has also been associated with a number of personal attributes. Aligning with the ABIM's essential humanistic qualities of integrity, respect, and compassion, Swick (2000) charged physicians to show evidence of core humanistic values, which, according to him, include honesty and integrity, caring and compassion, altruism and empathy, respect for others, and trustworthiness. Cohen (2007) related humanism to the manifestation of such attributes as altruism, duty, integrity, respect for others, and compassion. When providing tips on teaching and learning humanism, Cohen and Sherif (2014) conceptualized humanism as encompassing "honesty, integrity, caring, compassion, altruism, empathy, and respect for self, patients, peers, and other health care professionals" (p. 680). The Arnold Gold Foundation (2015) expects the humanistic healthcare professional to demonstrate the following attributes:

- **Integrity**: the congruence between expressed values and behavior
- **Excellence**: clinical expertise
- **Compassion**: the awareness and acknowledgement of the suffering of another and the desire to relieve it
- **Altruism**: the capacity to put the needs and interests of another before your own
- **Respect**: the regard for the autonomy and values of another person
- **Empathy**: the ability to put oneself in another's situation, e.g., physician as patient
- **Service**: the sharing of one's talent, time and resources with those in need; giving beyond what is required.

Medical administrators and educators generally agree that physicians are expected to demonstrate not only clinical competencies but also caring attitudes and behaviors. Many medical schools have designed and implemented a broad arrays of innovative educational programs intended to promote humanism in medical education (e.g., Branch et al. 2001; Branch et al. 2009; Ousager and Johannessen 2010; Stern et al. 2008). Although these intervention efforts usually receive positive feedback from the trainees and show short-term impact on gaining insight into patients' perspectives, a lack of long-term impact on the development of medical proficiency in providing humanistic care has been noted (Ousager and Johannessen 2010). One possible reason for the lack of long-term impact is the unavailability of appropriate assessment tools to capture the desired outcomes (Kuper 2006). The ABIM Subcommittee on Evaluation of Humanistic Qualities in the Internist (1983) recommended to the Board a need for continued research and the development of better methods for reliable, objective assessment of humanistic qualities in the internist. Instruments designed to assess humanistic patient care and supported by an evidence-based conceptual framework are urgently needed.

Based on the above literature review, we developed a conceptual framework to illustrate multifaceted dimensions of medical humanism and its relationships to patient-centeredness and professionalism, the two constructs most commonly associated with humanism with a level of confusion about their mutual connotations and individual scopes. As shown in Fig. 18.1, we conceptualize humanism as a

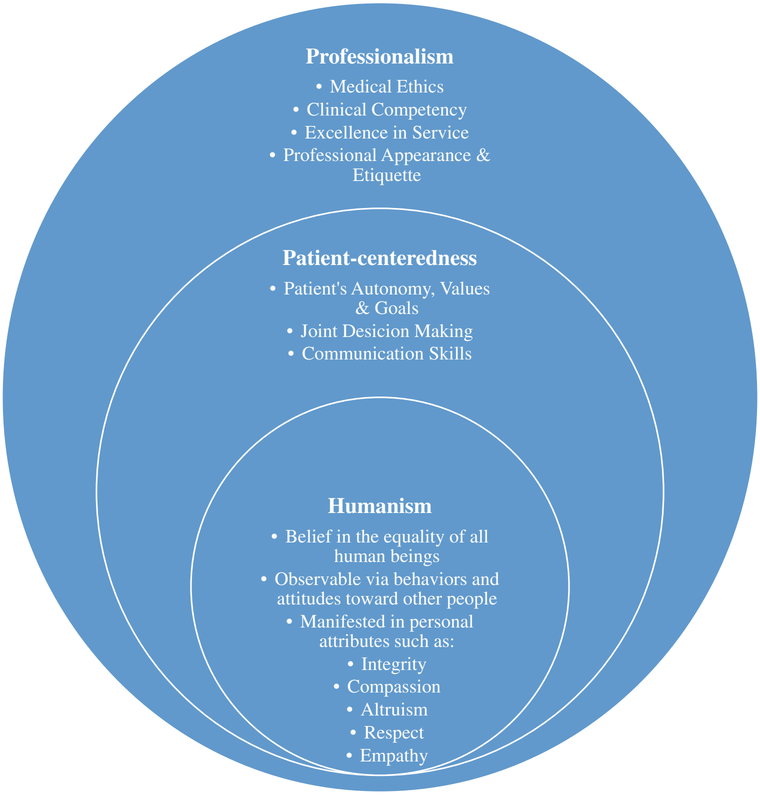

Professionalism
• Medical Ethics
• Clinical Competency
• Excellence in Service
• Professional Appearance & Etiquette
Patient-centeredness
• Patient's Autonomy, Values & Goals
• Joint Desicion Making
• Communication Skills
Humanism
• Belief in the equality of all human beings
• Observable via behaviors and attitudes toward other people
• Manifested in personal attributes such as:
• Integrity
• Compassion
• Altruism
• Respect
• Empathy

Fig. 18.1 Conceptual relationship among humanism, patient-centeredness, and professionalism in medicine

belief in the equality of all human beings, which manifests itself in such personal attributes as integrity, compassion, altruism, respect, and empathy and can be observed via an individual's behaviors and attitudes toward other people. Humanism is considered the core belief and value of medical professionals. When applied to clinical practice, medical professionals exhibit patient-centeredness in their communication and interaction with patients and their families by respecting their autonomy, values and goals as well as involving them in decision making for treatment options. The meaning of medical professionalism goes beyond humanistic attributes and patient-centered care to encompass additional dimensions such as medical ethics, clinical competency, excellence in service, and professional appearance and etiquette.

Informed by the need of sound tools for assessing humanistic patient care and guided by the conceptual framework described above, we conducted this study with a twofold aim: (1) to develop an observational scale to assess humanistic patient care; and (2) to conduct psychometric analyses to validate its use on medical students.

18.2 Methods

18.2.1 Instrument Development

Based on the conceptual framework, we drafted a 15-item, 5-point Likert scale with three items each in five subscales of Integrity, Compassion, Altruism, Respect, and Empathy (ICARE). The scale was subsequently used to assess a random sample of 50 students' videotaped performances in an Objective Structured Clinical Examination (OSCE) exam called Clinical Performance Examination (CPX). The CPX was administered to our medical students at the end of the third year. Students rotated through eight stations and spent fifteen minutes at each station to conduct a focused work-up on a trained standardized patient (SP). Students' clinical skills were rated by the SPs based on checklists developed and reviewed by faculty panels of the California Consortium for the Assessment of Clinical Competence, an inter-professional collaboration of healthcare professionals and medical educators from eight medical schools in the state of California. The ratings were summarized and reported to students by percentage of correct scores across all cases in five areas: history taking, physical examination, information sharing, physician-patient interaction, and patient-centered care. Students' videotaped performances in a depression case were used in the study and rated by three investigators, all experienced medical educators trained in educational or cognitive psychology. Our Institutional Review Board (IRB) approval was obtained before the study began.

A variation of the Delphi method (Linstone and Turoff 2011) was adopted in the process of training the three investigators to reach a consensus in the use of the rating scale and increase inter-rater reliability. Each investigator independently rated a sample of four to six videotaped CPX performances on the same case from a previous cohort before meeting in person. During the face-to-face meeting, they compared their ratings of every item for each student and discussed the rationale behind their ratings. They reached a consensus on how to use the scale and pertinent behavioral and attitudinal indicators for rating each item after viewing and rating the performance of a total of twenty students. The original scale was revised along the process through in-depth discussions. The training sessions also helped the investigators develop a comfortable confidence in conducting the task independently. The final version of the ICARE scale is included at the end of the chapter as an Appendix.

18.2.2 Instrument Validation

The revised scale was then used by the three investigators to assess the performances on the same depression case of a random sample of fifty students from another cohort, who were not included in the scale development stage.

To examine the reliability of the scale scores, Cronbach's alpha (Cronbach 1951) estimating internal consistency reliability and the intraclass correlation (ICC) (Shrout and Fleiss 1979) examining inter-rater reliability were calculated. In addition, we conducted the generalizability (G) study (Shavelson and Webb 1991) to examine reproducibility reliability and cross-check internal consistency and inter-rater reliability.

To examine the construct validity of the instrument, we conducted exploratory factor analysis to investigate the underlying structure of the scale. We used principal axis factoring and direct oblimin rotation as we expected the latent constructs to be correlated. In addition, Pearson Product Moment correlation coefficients were calculated between the scores on four external criterion measures and ICARE ratings to examine the criterion-related validity. The four criterion measures were (1) the CPX checklist on the patient–physician interaction (PPI) items; (2) the CPX checklist on the patient-centered care (PCC) items; (3) the Jefferson Scale of Physician Empathy (JSPE) (Hojat et al. 2001); and (4) the Patient–Practitioner Orientation Scale (PPOS) (Krupat et al. 1996) with Sharing and Caring subscales. For the CPX-PPI and CPX-PCC checklists, we used scores on the depression case only and the mean scores across all eight stations.

We also conducted t-tests to compare the differences between mean ICARE subscale scores and the differences between male and female student scores across all the measures. The analyses were conducted using IBM SPSS Statistics 20 (IBM Corp., Armonk, NY) and GENOVA (Brennan 2001).

18.3 Results

The random sample of 50 students represented one third of the class. Male (48 %) and female students were about equally represented in the sample. Mean and standard deviation scores by gender on all measures included in the study are shown in Table 18.1. The mean score on the Respect subscale was significantly ($p < 0.05$) higher than those of all the other subscales. All the subscale mean scores were also significantly ($p < 0.05$) different from each other except those between the Empathy and Compassion subscales. There was no significant gender difference in any measures included in the study except the JSPE scale where women scored significantly ($p < 0.05$) higher than men.

18.3.1 Reliability

The internal consistency reliability estimated by Cronbach's alpha for the entire 15-item ICARE scale was 0.95, with the alpha for the subscales ranging from 0.78 (Respect), 0.87 (Integrity), 0.89 (Altruism), to 0.92 (Empathy and Compassion, respectively). The ICC coefficient for the whole scale was 0.64, and the subscale coefficients were 0.41 (Respect), 0.51 (Altruism), 0.59 (Compassion), 0.62 (Integrity), and 0.67 (Empathy).

Table 18.2 shows the estimated variance components from the G study. The variance component associated with Person (σ^2_{p}, 57 % of the total variation) was fairly large compared to the other components. This indicates that, averaged over raters and items, students in the study differed in their humanistic care performance. The second largest variance component was associated with the interaction between Person and Raters (σ^2_{pr}), representing 29 % of the total variation. This shows that the scores students received were dependent on the rater. The raters, however, tended to independently rate very consistently, as the percent of total variance accounted for by the Rater component (σ^2_{r}) was 0. The non-negligible 5 % of the total variance associated with Subscale (σ^2_{s}) shows that rating of the subscales varied somewhat in level of difficulty. The relatively small Residual component ($\sigma^2_{\mathrm{pri:s,e}}$, 2.8 %) reflects that the majority of important sources of variance had been accounted for by the varying relative standings of persons across subscales and items, and other related sources of error. Finally, a G coefficient ($E\rho^2$) of 0.61 and a Dependability coefficient (ϕ) of 0.57 indicate, respectively, that the ICARE scale is moderately good at ranking people and moderately dependable at identifying students proficient in humanistic care.

Table 18.1 Mean scores of all measures by gender

Measure	Male N = 24		Female N = 26		Total N = 50	
	Mean	SD	Mean	SD	Mean	SD
ICARE_Total[a]	3.83	0.57	3.66	0.53	3.74	0.55
ICARE_Integrity[a]	4.06	0.61	3.80	0.52	3.92	0.57
ICARE_Compassion[a]	3.69	0.77	3.64	0.66	3.66	0.71
ICARE_Altruism[a]	3.51	0.76	3.37	0.67	3.44	0.71
ICARE_Respect[a]	4.25	0.58	3.92	0.44	4.08	0.53
ICARE_Empathy[a]	3.62	0.58	3.58	0.69	3.60	0.63
PPOS_Total[b]	4.38	0.45	4.49	0.46	4.44	0.45
PPOS_Sharing[b]	4.53	0.52	4.71	0.59	4.63	0.56
PPOS_Caring[b]	4.22	0.47	4.27	0.49	4.25	0.48
JSPE[c]	4.80	0.56	5.14	0.43	4.98	0.52
CPX-PPI_Total[d]	70.18	7.11	73.65	6.03	71.99	6.73
CPX-PCC_Total[d]	70.87	9.08	73.76	7.12	72.37	8.17
CPX-PPI_Depression[d]	65.48	13.40	71.43	15.12	68.57	14.49
CPX-PCC_Depression[d]	71.30	9.22	76.50	13.01	74.00	11.54

There was no significant gender difference in any measures except the JSPE scale where women scored significantly ($p < 0.05$) higher than men

[a] gender (in the table title)

[b] ICARE

[c] PPOS

[d] JSPE

[e] CPX

[f] all ICARE total scale and subscales (i.e., ICARE_Total, ICARE_Integrity, ICARE_Compassion, ICARE_Altruism, ICARE_Respect, ICARE_Empathy)

Table 18.2 Estimated variance components in the generalizability (G) study

Source of variation	Degree of freedom (df)	Variance component	Estimated variance component	Percent of total variance
Person *(p)*	19	σ^2_{p}	0.11762	56.55
Rater (*r*)	2	σ^2_{r}	−0.00760	0
Subscale *(s)*	4	σ^2_{s}	0.01047	5.03
Item (*i:s*)	10	$\sigma^2_{\mathrm{i:s}}$	0.00319	1.53
Person-by-rater (*pr)*	38	σ^2_{pr}	0.06080	29.23
Person-by-subscale (*ps)*	76	σ^2_{ps}	0.00457	2.20
Person-by-item (*pi:s)*	190	$\sigma^2_{\mathrm{pi:s}}$	0.00096	0.46
Rater-by-subscale (*rs)*	8	σ^2_{rs}	0.00167	0.80
Rater-by-item (*ri:s)*	20	$\sigma^2_{\mathrm{ri:s}}$	0.00044	0.21
Person-by-rater-by-subscale (*prs)*	152	σ^2_{prs}	0.00238	1.14
Residual (*pri:s,e)*	380	$\sigma^2_{\mathrm{pri:s,e}}$	0.00591	2.84
Generalizability (Eρ^2)				0.61
Dependability coefficient (ϕ)				0.57

Table 18.3 Exploratory factor analysis of the ICARE scale: Factor loadings and inter-factor correlation after rotation

Subscale	Item	Factor 1	Factor 2
Integrity (I)	Consistently expresses a genuine concern for patient via both verbal and non-verbal behaviors	0.63	**0.65**
	Shows the quality of being honest	0.65	**0.82**
	Adheres to a strict moral or ethic standard	0.63	**0.79**
Compassion (C)	Recognizes the suffering experienced by patient	**0.84**	0.57
	Expresses a desire to alleviate patient's suffering	**0.96**	0.44
	Exhibits a determination to provide the best care to patient	**0.89**	0.44
Altruism (A)	Puts patient's needs and interests first	**0.85**	0.56
	Exhibits selfless concern for the welfare of patient	**0.89**	0.40
	Demonstrates a willingness to sacrifice oneself to provide services	**0.75**	0.56
Respect (R)	Treats patient with courtesy	0.29	**0.69**
	Gives patient enough time to respond	0.41	**0.80**
	Asks patient for his/her opinions about diagnosis and treatment plan	0.49	**0.67**
Empathy (E)	Demonstrates understanding of patient's feelings and experiences	**0.85**	0.62
	Is able to relate or refer to patient's feelings and experiences	**0.82**	0.42
	Shows shared feelings with patient	**0.86**	0.69
Inter-factor correlations	Factor 1		0.55

Extraction method: Principal Axis Factoring; Rotation method: Direct oblimin

18.3.2 Validity

The exploratory factor analysis revealed two factors with an eigenvalue greater than 1. The two factors accounted for 73 % of the total variance. Table 18.3 shows the factor loadings of each item on the two extracted factors and inter-factor correlations after rotation. All the items of the Empathy, Compassion, and Altruism subscales loaded heavily on Factor 1. All the items of the Respect and Integrity subscales loaded heavily on Factor 2. The two factors were hence labeled Compassionate Care (Factor 1) and Professional Care (Factor 2). The inter-factor correlation coefficient showed a moderate correlation (0.55) between Compassionate Care and Professional Care.

The correlations between the total and subscale mean scores of the ICARE scale and the mean scores of the four external criterion measures are presented in Table 18.4. Most of the correlations between the ICARE scale and subscales and the criterion measures were at either a low positive level ($r = 0.25$ or below) or a level close to zero, indicating no meaningful association between humanistic care assessed by the ICARE and patient-centeredness and empathy assessed by the criterion measures.

Table 18.4 Correlations of the Scores on the ICARE Scale with the Scores on the External Criterion Measures

	ICARE_ Total	ICARE_ Integrity	ICARE_ Compassion	ICARE_ Altruism	ICARE_ Respect	ICARE_ Empathy	PPOS_ Total	PPOS_ Sharing	PPOS_ Caring	JSPE	CPX PPI_Total	CPX PCC_Total	CPX PPI_Depression
ICARE_Total													
ICARE_Integrity	0.86**												
ICARE_Compassion	0.92**	0.70**											
ICARE_Altruism	0.91**	0.72**	0.87**										
ICARE_Respect	0.70**	0.67**	0.43**	0.47**									
ICARE_Empathy	0.91**	0.66**	0.88**	0.79**	0.58**								
PPOS_Total	0.08	0.04	−0.01	0.12	0.20	0.04							
PPOS_Sharing	0.10	0.02	0.06	0.16	0.11	0.10	0.90**						
PPOS_Caring	0.04	0.05	−0.08	0.03	0.25	−0.04	0.86**	0.54**					
JSPE	0.14	0.07	−0.00	0.14	0.25	0.18	0.58**	0.54**	0.47**				
CPX PPI_Total	0.20	0.14	0.22	0.20	0.01	0.25	0.33*	0.41**	0.15	0.40**			
CPX PCC_Total	0.12	0.02	0.23	0.19	−0.17	0.18	0.13	0.34*	−0.16	0.22	0.83**		
CPX PPI_Depression	0.16	0.09	0.18	0.10	0.12	0.20	0.14	0.05	0.21	0.03	0.47**	0.22	
CPX PCC_Depression	0.15	0.09	0.17	0.10	0.09	0.17	0.11	0.04	0.17	0.09	0.49**	0.27	0.94**

Pearson Product Moment correlation coefficients were reported
*Correlation is significant at the 0.05 level (2-tailed)
**Correlation is significant at the 0.01 level (2-tailed)

18.4 Discussion

A review of the literature in medical humanism and its relationship to professionalism and patient-centeredness directed us to develop a conceptual framework delineating the conceptual relationships among the three commonly associated constructs. The framework conceptualizes humanism as a core belief and value of medical professionalism, manifested in five personal attributes (integrity, compassion, altruism, respect, and empathy) and observable in patient-centered care behaviors and attitudes. Based on this framework, we developed a short, 15-item ICARE observational scale to assess medical students' performance in humanistic patient care in a simulated clinical setting. A follow-up validation process provided support for the reliability and construct validity of the scale scores. The high internal consistency reliability of the whole scale and the five subscales indicates that the 15 items were closely related and represented items in the universe measuring the same underlying constructs. The sound reliability is also shown in the G study by the relatively high proportion (57 %) of total variance explained by the Person component, indicating the scale is capable of capturing systemic individual differences in humanistic performance. The G coefficient of 0.61 further supports that the instrument is good at distinguishing people based on performance. The moderate intraclass correlation coefficients, together with a relatively high proportion (29 %) of total variance in the interaction between raters and students, indicate that better calibration of raters is needed to more accurately reflect students' performances in humanistic patient care and further improve score reliability. Nonetheless, a modified Delphi method used in the scale development process was helpful in establishing strong internal consistency reliability and a moderate inter-rater reliability. The process facilitated the development of a set of behavioral and attitudinal indicators for assessing students' performance in each domain. The consensus reached through this method can also be used in future faculty development to train additional raters.

Exploratory factor analysis identified two latent constructs of the ICARE scale, labeled Compassionate Care and Professional Care. The items clustered under the Compassionate Care factor indicate warm regards toward patients and a person-orientated approach to patient care. In contrast, the items clustered under the Professional Care factor entail conduct appropriate to medical professionals. The findings facilitate the understanding of the latent structure of the ICARE scale.

The low correlations of the ICARE scale and subscale scores with those of the criterion measures may be explained by one or a combination of the following reasons: (1) students' humanistic performance in a simulated clinical setting differed from their self-perceived patient-centeredness and empathy; (2) faculty raters' assessment of student performance deviated from students' own self-perceptions and standardized patients' assessments; and (3) slightly higher correlations in the scores of PPOS and JSPE scales with the Respect subscale, as well as in CPX-PPI and PCC scores with the Compassion and Empathy subscales, suggest that the domains these measures assess intersect, but each measure also assesses certain unique domains. The criterion-related validity of the ICARE scale remains to be further investigated.

A lack of findings suggesting gender differences in humanistic care indicates that both male and female students are capable of providing compassionate and professional care to patients. Although empathy was one of the areas assessed by the ICARE scale, which has been documented in the literature (e.g., Krupat et al. 2000) to show a stronger inclination among women than men, our instrument assesses multifaceted dimensions of humanistic care in which men and women performed equally.

The finding that the mean scores of the Respect and Integrity subscales were significantly higher than the mean scores of all the other subscales is also worth noting. As patient-centeredness and professionalism have received broad attention in medical education in recent years, students have been well trained to show respectfulness and integrity in clinical settings, which leads to the higher subscale scores found in this study. Compassionate and altruistic attitudes and behaviors in patient care remain in need of greater emphasis in medicine curriculum. Perhaps clinician educators would like to consider emphasizing compassion and altruism in courses such as Doctoring and Preceptorship as well as all core clerkship rotations. Medical administrators would consider launching faculty development programs focusing on medical humanism so that proper role models may be observed by trainees of all levels.

Our study has several limitations. First, it was conducted in one institution. Other institutions embracing different curricular designs and student populations may produce different results. Second, only one OSCE case, depression, was used in the study. While the students' humanistic attitudes and behaviors toward a depressed patient might be easily observed and assessed by using the ICARE scale, the scale may not work equally well in assessing the care provided to patients with other conditions. The applicability of the scale to assessing patient care in general needs to be further investigated. Third, only a small sample of students participated in this study. Although the generalizability coefficient was moderately high, replication of the study by using a large sample across cohorts or institutions is warranted. Finally, the scale provides means to collect quantitative data only. To better and more comprehensively understand trainees' humanistic inclination, a qualitative method (e.g., semi-structured interview, reflective write-ups) or even a 360 degree approach (multi-method, multi-rater, and multi-setting) may deserve consideration and exploration.

18.5 Conclusion

This chapter presents a conceptual framework to clarify the connotations of medical humanism and its relationships to professionalism and patient-centeredness, and a short observational instrument for assessing humanistic patient care. As healthcare professionals continue the efforts in advancing their professional craftsmanship, the conceptual framework is developed in hopes that it will provide guidelines for healthcare educators in reforming curriculum toward training humanistic practitioners who can deliver patient-centered professional care. Although the study demonstrated sound psychometric properties of the ICARE scale in the assessment of humanistic care attitude and behavior, further refinement of the tool may be necessary. Ensuing steps in advancing the scale may include: (1) recruiting clinician educators, rather than the medical educators involved in the scale development stage who were not trained in clinical practice, to review and refine the instrument for better clinical application; (2) cross-validating the refined instrument by applying it to assess trainees of different levels in multiple OSCE stations as well as real clinical settings; and (3) training SPs and clinician educators in using the instrument for assessment purposes.

Questions for Reflection

- Where and how in healthcare curricula is it best to emphasize humanistic patient care?
- How can the assessment of humanistic patient care in authentic or simulated clinical settings be implemented in a reliable and feasible way?
- How are clinical role models in medical humanism best trained and sustained?
- What mechanisms are needed to promote medical humanism in the current culture of medicine?

Appendix

ICARE scale for assessing humanistic patient care

Dimension/element/indicator	Poor	Below average	Average	Above average	Excellent	NA
Integrity						
• Expresses consistently a genuine concern of patient via both verbal and non-verbal behaviors – Demonstrates an overall demeanor as a genuinely caring person	1	2	3	4	5	NA
• Shows the quality of being honest – Provides true answers, based on own knowledge level, to patient	1	2	3	4	5	NA

(continued)

(continued)

Dimension/element/indicator	Poor	Below average	Average	Above average	Excellent	NA
– Admits own limits in treating patient						
• Adheres to a strict moral or ethical standard – Suggests the most righteous treatment options for patient	1	2	3	4	5	NA
Compassion						
• Recognizes the sufferings experienced by patient – Verbally points out patient's suffering – Uses body language (handing over tissues, stopping notes taking) to show recognition	1	2	3	4	5	NA
• Expresses a desire to alleviate patient's suffering – Uses body language (moving closer to patient, properly patting patient) to comfort patient – Says something like "We will bring you back to your normal life."	1	2	3	4	5	NA
• Exhibits a determination of providing the best care to patient – Make efforts to go beyond the standard care by comparing and contrasting different treatment options for patient	1	2	3	4	5	NA
Altruism						
• Puts patient's needs and interests first – Focuses on listening or responding to patient, instead of taking notes, when patient communicating emotional experiences	1	2	3	4	5	NA
• Exhibits selfless concern for the welfare of patient – Continues focusing on patient's wellbeing even when time is up	1	2	3	4	5	NA

(continued)

(continued)

Dimension/element/indicator	Poor	Below average	Average	Above average	Excellent	NA
• Demonstrates a willingness to sacrifice oneself to provide services – Makes self available beyond regular hours for patient	1	2	3	4	5	NA
Respect						
• Treats patient with courtesy – Greets patient properly – Maintains eye contact – Uses proper verbal tone – Pays attention to patient's comfort level when interacting with patient – Praises patient for good health behaviors/habits – Completes the interview with a proper closure	1	2	3	4	5	NA
• Gives patient enough time to respond – Does not improperly interrupt patient's communications	1	2	3	4	5	NA
• Asks patient for his/her opinions about diagnosis and treatment plan – Encourages patient to ask questions – Encourages patient to talk about difficult issues – Checks patient's preference for treatment options	1	2	3	4	5	NA
Empathy						
• Demonstrates understanding of patient's feelings and experiences – Nods to show understanding or acknowledgement – Repeats or rephrases what patient said – Follows up properly with what patient said	1	2	3	4	5	NA
• Be able to relate or refer to patient's experiences	1	2	3	4	5	NA

(continued)

(continued)

Dimension/element/indicator	Poor	Below average	Average	Above average	Excellent	NA
– Shares similar personal experiences – Acknowledges patient's experiences as something commonly shared by general public						
• Shows shared feelings with patient – Exhibits same feelings as patient's (smile/laugh, tears, cracking voice)	1	2	3	4	5	NA

References

American Board of Internal Medicine Subcommittee on Evaluation of Humanistic Qualities in the Internist. (1983). Evaluation of Humanistic Qualities in the Internist. *Annuals of Internal Medicine, 99*, 720–724.

American Board of Internal Medicine. (1995). *Project professionalism*. Retrieved July 15, 2015, from http://www.abimfoundation.org/~/media/Foundation/Professionalism/Project%20professionalism.ashx?la=en

Arnold Gold Foundation Gold Humanism Honor Society. (2005). *A force for humanism in medicine*. Englewood Cliffs, NJ: The Arnold P. Gold Foundation.

Arnold Gold Foundation. (2015). *Frequently asked questions*. Retrieved July 15, 2015, from http://humanism-in-medicine.org/about-us/faqs/

Bever, C. T, Jr., Franklin, G. M., Kaufman, J. M., & Esper, G. J. (2010). Role of professionalism in improving the patient-centeredness, timeliness, and equity of neurological care. *Archives of Neurology, 67*, 1386–1390.

Branch, W. T, Jr., Kern, D., Haidet, P. M., Weissmann, P. F., Gracey, C. F., Mitchell, G. A., et al. (2001). Teaching the human dimensions of care in clinical settings. *Journal of American Medical Association, 286*, 1067–1074.

Branch, W. T, Jr., Frankel, R., Gracey, C. F., Haidet, P. M., Weissmann, P. F., Cantey, P., et al. (2009). A good clinician and a caring person: Longitudinal faculty development and the enhancement of the human dimensions of care. *Academic Medicine, 84*, 117–126.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cohen, J. J. (2007). Linking professionalism to humanism: What it means, why it matters. *Academic Medicine, 82*, 1029–1032.

Cohen, L. G., & Sherif, Y. A. (2014). Twelve tips on teaching and learning humanism in medical education. *Medical Teacher, 36*, 680–684.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Davis, K., Schoenbaum, S. C., & Audet, A. (2005). A 2020 vision of patient-centered primary care. *Journal of General Internal Medicine, 20*, 953–957.

Gold, A., & Gold, S. (2006). Humanism in medicine from the perspective of the Arnold Gold Foundation: Challenges to maintaining the care in health care. *Journal of Child Neurology, 21*, 546–549.

Goldberg, J. L. (2008). Humanism or professionalism? The White Coat Ceremony and medical education. *Academic Medicine, 83*, 715–722.

Hartzband, P., & Groopman, J. (2009). Keeping the patient in the equation—Humanism and health care reform. *New England Journal of Medicine, 361*, 554–555.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J. M., Gonnella, J. S., Erdmann, J. B., et al. (2001). The Jefferson scale of physician empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement, 61*, 349–365.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century* (Vol. 6). Washington, DC: National Academy Press.

Krupat, E., Putnam, S. M., & Yeager, C. (1996). The fit between doctors and patients: can it be measured? *Journal of General Internal Medicine, 11*, 134.

Krupat, E., Rosenkranz, S. L., Yeager, C. M., Barnard, K., Putnam, S. M., & Inui, T. S. (2000). The practice orientations of physicians and patients: The effect of doctor-patient congruence on satisfaction. *Patient Education and Counseling, 39*, 49–59.

Kuper, A. (2006). Literature and medicine: A problem of assessment. *Academic Medicine, 81*, S128–S137.

Laine, C., & Davidoff, F. (1996). Patient-centered medicine: A professional evolution. *Journal of American Medical Association, 275*, 152–156.

Lévesque, M. C., Hovey, R. B., & Bedos, C. (2013). Advancing patient-centered care through transformative educational leadership: A critical review of health care professional preparation for patient-centered care. *Journal of Healthcare Leadership, 5*, 35–46.

Liaison Committee on Medical Education. (2011). Functions and structure of a medical school: Standards for accreditation of medical education programs leading to the M.D. degree. Washington, D.C: Liaison Committee on Medical Education.

Linstone, H. A., & Turoff, M. (Eds.). (2011). The Delphi method: Techniques and applications. Retrieved February 2, 2011, from http://is.njit.edu/pubs/delphibook/delphibook.pdf

Mead, N., & Bower, P. (2000). Patient-centredness: A conceptual framework and review of the empirical literature. *Social Science and Medicine, 51*, 1087–1110.

Miller, S., & Schmidt, H. (1999). The habit of humanism: A framework for making humanistic are a reflexive clinical skill. *Academic Medicine, 74*, 800–803.

Ousager, J., & Johannessen, H. (2010). Humanities in undergraduate medical education: A literature review. *Academic Medicine, 85*, 988–998.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Stern, D. T., Cohen, J. J., Bruder, A., Packer, B., & Sole, A. (2008). Teaching humanism. *Perspectives in Biology and Medicine, 51*, 495–507.

Stern, D. T., & Papadakis, M. (2006). The developing physician—becoming a professional. *New England Journal of Medicine, 355*, 1794–1799.

Swick, H. M. (2000). Toward a normative definition of medical professionalism. *Academic Medicine, 75*, 612–616.

Swick, H. M. (2007). Viewpoint: Professionalism and humanism beyond the academic health center. *Academic Medicine, 82*, 1022–1028.

Chapter 19
Evaluating the Paradigm Shift from Time-Based Toward Competency-Based Medical Education: Implications for Curriculum and Assessment

Yoon Soo Park, Brian D. Hodges and Ara Tekian

Abstract In the early twentieth century, most curricula were based on a concept of fixed time. Students who successfully completed a program were judged to be competent. However, a paradigm shift toward competency-based education occurred at the end of the twentieth century, allowing only students who are judged "competent" to move forward in a professional school curriculum. There are significant implications to this paradigm shift, particularly for curricular design, performance assessment, faculty development, and resources. Educators may find challenges addressing individual learning differences—some students are able to progress easily in some subject areas, while some may continue to struggle. Learners can also progress at different rates in competency-based education programs. While it is relatively easy to develop competencies in areas of knowledge and skill, it is more difficult to define milestone assessments in areas such as reasoning and judgment, and to assess complex professional behaviors. The promise in competency-based education is to graduate professionals who are better adapted for the needs of complex and rapidly changing systems. Yet, implementing competency-based curricula raise important questions not only in instruction, but also in the assessment of competencies and outcomes. This chapter synthesizes the existing literature and perspectives that support and critique competency-based education, identifies pressing challenges for educators, and speculates on the future of this still emerging paradigm.

Y.S. Park (✉)
Department of Medical Education, University of Illinois at Chicago,
Chicago, IL, USA
e-mail: yspark2@uic.edu

B.D. Hodges
University of Toronto, Toronto, Canada

A. Tekian
University of Illinois at Chicago, Chicago, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_19

Takeaways

- Traditional curricular structure in medical education has been defined through a *time-based model*, corresponding to fixed time spent in training.
- Scholars and practitioners have called for a move toward *competency-based model* that ensures the achievement of competencies and entrustment of skills as outcomes, driven by societal expectations and public accountability.
- Implementing a competency-based model accompanies challenges in instruction and assessment, faculty development, and allocation of resources.
- Designing a multifaceted and learner-centered workplace-based assessment in a competency-based setting requires overcoming traditional notions of psychometrics and interpretation of validity. Narrative assessments and subjective judgment models may offer new insights to address these issues.

19.1 Introduction

The movement toward reforming medical education curriculum has rapidly reshaped the instruction and assessment of learners entering the health professions workforce. Traditionally, the signature of medical education curriculum has been defined through a “structure/process” format, as outlined in Abraham Flexner’s legacy—*Medical Education in the United States and Canada. A Report to the Carnegie Foundation for the Advancement of Teaching* (Flexner 1910). While Flexner’s model introduced a scientific and evidence-based approach to teaching medicine, it is perhaps most noted for the creation of a binary structure to medical education curriculum in which students are taught basic sciences followed by clinical training. By the turn of the century, traditional curriculum in medical education was being challenged, targeting the need to emphasize competence development and entrustment to better meet public demands (Englander et al. 2015a; ten Cate and Scheele 2007). These challenges are in fact, not so foreign to Flexner’s philosophy; Flexner also advocated for other issues in medical training, including societal expectations of the medical profession (Whitehead 2010).

The traditional curricular structure in medical education is commonly associated with a *time-based model*, where training corresponds to a fixed time spent in medical school or graduate medical training (Hodges 2010; Tekian et al. 2015; Gruppen et al. 2012). During the past two decades, the time-based model has received increasing attention and criticism from the health professions education community in North American and around the world. Calls for a shift toward *competency-based model* are receiving greater attention. Scholars, practitioners, and policymakers have voiced concerns over the fixed time approach that is currently dominant in the education system. For example, in the United States, concerns about lack of sufficient training and inadequate readiness to practice following residency have been reported (Crosson et al. 2011). Ensuring the preparedness of newly trained health professionals with the knowledge and skills required for everyday practice is part of an ongoing movement documented in a recent report by the Institute of Medicine (IOM 2014). Moreover, it is within these initiatives that the Accreditation Council for Graduate Medical Education (ACGME) launched the Next Accreditation System (NAS), whereby resident progress is tracked through the achievement of developmental milestones that target each of the six ACGME core competencies (Nasca et al. 2012). Similar movements toward competency-based models are emerging throughout the world. It is at this tipping point that implications for teaching and assessing of specific competencies require further investigation.

While the concept of "competence" has received increasing attention from the field, a consensus definition of the term and a collective rationale for movement toward competency-based medical education (CBME) curriculum has received varying reactions. In this regard, this chapter aims to review the existing literature on the paradigm shift from time-based toward CBME curriculum; strengths and challenges of both models are identified. Moreover, challenges in implementing CBME and relevant assessment frameworks are discussed, including potential barriers faced with current conceptual notions of validity and assessment standards. Implications for the CBME movement in relation to existing assessment guidelines and notions of composite validity (AERA, APA and NCME 2014) and subjective judgment models are also discussed.

19.2 Two Models of Curriculum in Medical Education: Time and Competency

The current literature on competence development identifies two curricular models: (1) time-based models and (2) competency-based models. This section describes the two models and the current trend toward CBME.

19.2.1 Time-Based Model

Tea-steeping model. During the early twentieth century, Flexner's recommendation for scientific and evidence-based training signaled an alignment between clinical and basic sciences in medical education. This subsequently prompted medical schools to join or affirm relationships with universities (Starr 1982; Gidney and Millar 1994). However, universities at the time already conceptualized competence as a time-based tradition, using fixed length programs of study with the mastery of knowledge and replication of facts and information as the central focus of assessments. This naturally led to a fixed-time model for medical education, which came to be divided into the foundational, basic sciences phase and a clinical phase. Within the time-based model, undergraduate medical education curriculum was characterized by a 2 + 2 structure in most North American programs or a 4 + 2 structure for programs in Europe for the basic sciences and for clinical years, respectively. The concept of fixed time also extended to graduate medical education, in which trainees graduated after an established duration of training time. Moreover, the notion of "rotations" also originated from universities that were affiliated with clinics and hospitals. As such, the organization and structure of time-based model in medical education curriculum can be seen as being rooted in the university's existing tradition (Hodges 2010).

Based on an assumed sufficiency of time as a condition for competence in medicine, Hodges (2010) coined the "tea-steeping" metaphor for the time-based model. Here, students are likened to tea leaves that are soaked for a fixed duration in hot water (i.e., medical school). In this view, changes to admissions policy are analogous to changing the type of tea, while altering or changing the curriculum or the school environment is simply changing the nature of water that is used to brew the tea.

Challenges to the time-based model. Two fundamental issues arise with the fixed, time-based model. First, the assessment structure of competence and proper entrustment of skills can be challenging, as it lacks the flexibility to meet individual progression toward competence. In time-based models, time is so central to completion that learners may graduate without having properly demonstrated mastery in required competencies. To address this limitation, a system to monitor and provide ongoing feedback for learners to achieve competence within the fixed time interval is needed. However, courses in time-based models are structured into predetermined intervals, making ongoing, individualized feedback structurally difficult; that is, monitoring and providing individualized feedback to learners in time-based models—particularly on achieving competencies—can be limited due to the inherent condition of graduating learners within a fixed time period.

While summative assessments are common in time-based curricular models, they do not necessarily provide information about whether learners have acquired the competencies to *perform* at work. Implementing workplace-based assessments (WBAs) that allow continuous measurement of skills in the clinical workplace can be difficult in a time-based environment, because of a lack of alignment between instruction in time-based modules and assessment of overall educational outcomes. Most curricular structure is oriented to completing a time-based block rather than aiming toward overarching educational outcomes. Furthermore, individual rates at which learners acquire necessary competencies may not be measured or tracked within the time-based model. This can also deter promoting self-regulated and flexible learning environments for learners (Gruppen et al. 2012). Because learners progress at different rates, some may achieve the required competencies prior to graduation. A one-size-fits all assessment policy rooted in time-based curricula may not be an adequate model for learning in the twenty-first century.

Second, changes or reforms are difficult to implement in a time-based model; altering curriculum often means only making minor adjustments to accommodate change, rather than a wholesale reform driving the curriculum. Within the traditional time-based curricula, proposals to modify the curriculum often take an additive approach, whereby new topics such as ethics or communications are simply added, rather than streamlining or removing existing content. Therefore, with learners' schedules already fully committed, modifications to curricula may only result in modest changes.

Given these structural issues in the time-based model, ensuring competence development can be problematic. Moreover, in an era when public accountability and societal expectations drive the need for evidence of competence development, the imperative for curricular reform becomes more pressing. Although the time-based model has adequately served the needs of health professionals since Flexner's era, these concerns have led to a movement toward CBME.

19.2.2 Competency-Based Models

Limitations of the time-based model have fueled discussions about adopting new, non-time-based, competency-based models. In this section, the movement and motivation toward competency-based models are introduced, including defining competence from a historic perspective. Factors that constituted competence of a health professional over a century ago during Flexner's era were quite different from the complex competency frameworks used today; a historic understanding of the evolution and conceptualization of competence will provide background for movement toward CBME.

Transition to "entrustment" as potential outcomes in CBME. The defining elements of competence during the early twentieth century, when the time-based model was being formulated, were knowledge or knowledge-based clinical performance. However, toward the end of the twentieth century, notions of competence evolved. During the latter twentieth century, competence frameworks extended beyond knowledge to include quality of patient care, interpersonal and communication skills, professionalism, and teamwork, among others. Moreover, with the introduction of Miller's Pyramid (Miller 1990), assessment of competence was regarded to include not only what a learner *knows*, but also what he or she *does*. The transition from conceptualizing competence as knowledge to conceiving it as a set of "entrustable" skills was an important shift between views of competence during Flexner's era and today.

Competency-based models of medical education curriculum are closely related to *outcome-based models* that were popular in general education in the mid twentieth century. Within outcome-based models, measurable competencies are defined, for which learners are trained; these measurable competencies can, in theory, lead to the creation of individualized, outcome-based curricula. In competency-based models, the outcomes originate from the needs of the community; such community-based needs inform the desired competencies and outcomes of training, which in turn lead to constructing the curriculum and assessment (Tekian et al. 2015). Many competency-based models require the identification of *entrustable professional activities* (EPAs; ten Cate and Scheele 2007; Englander et al. 2015b), which mark measurable outcomes or competencies in CBME. It is important to emphasize that within such competency-based models, competence is based on *what learners do*—the highest level of Miller's Pyramid. Placing emphasis on actual practice abilities (including "entrustable professional acts") differs significantly from previous notions of assessments used to determine what learners *know how* to do or even *show how* to do—the latter two functioning as proxies for what learners actually *do* in practice.

19.2.2.1 Recommendations from North America that Motivate Medical Education Reform

The movement toward CBME was fueled by reports released from the United States and Canada. In 2010, both the United States and Canada released recommendations for reforming medical education. These reports were created to advance the state of medical education, noting that in many ways, there were modest changed in the 100 years since Flexner's 1910 report. This section summarizes the North American recommendations for medical education, emphasizing principles of CBME.

Recommendations from the United States and Canada. In the United States, four recommendations toward medical education reform were reported in *Educating Physicians: A Call for Reform of Medical School and Residency* (Cooke et al. 2010). Table 19.1 summarizes these recommendations: (1) standardizing outcomes, (2) integrating, (3) fostering habits of inquiry, and (4) forming an identity. Standardizing outcomes refers to both learning and practice outcomes; the report emphasizes identifying competencies and milestones and the use of multiple forms of assessments. Integration refers to linking knowledge and experience, while engaging in different forms of reasoning (analytic, pattern recognition, creative, and adaptive). For habits of inquiry and improvement, the report recommends developing expertise through deliberate practice and feedback, while engaging in communities of inquiry and practice. Finally, for identity formation, the report recommends a commitment to values, participation in community practice, observing role models, and feedback. A similar set of recommendations was released in Canada by the Association of Faculties of Medicine of Canada (AFMC): *The Future of Medical Education in Canada (FMEC): A Collective Vision for MD Education* (AFMC 2010) for medical students and *A Collective Vision for Postgraduate Medical Education in Canada* (AFMC 2012) for postgraduates (see Table 19.1).

Common themes across the United States and Canada. A common theme across both undergraduate and postgraduate medical education from the United States and Canada is a call for outcomes and competency-based curriculum.

Table 19.1 Recommendations for medical education in the United States and Canada

United States[a] (Cooke et al. 2010)	Canada (AFMC 2010, 2012)	
	Undergraduate[b]	Postgraduate[c]
1. Standardize on outcomes that can allow flexibility in learning 2. Integrate knowledge and experience 3. Foster habits of inquiry and improvement that focus on excellence 4. Create professional identity that carry professional values and dispositions	1. Address individual and community needs 2. Enhance admissions processes 3. Build on the scientific basis of medicine 4. Promote prevention and public health 5. Address the hidden curriculum 6. Diversify learning contexts 7. Value generalism 8. Advance inter- and intra-professional practice 9. Adopt a competency-based and flexible approach 10. Foster medical leadership	1. Right mix, distribution and numbers of physicians 2. Diverse learning and work environments 3. A positive and supportive environment 4. Competency based curricula 5. Transitions along the medical educational continuum 6. Effective assessments systems 7. Support clinical teachers 8. Foster leadership development 9. Collaborative governance in PGME 10. Align accreditation standards

Note
[a] Cooke et al. (2010)
[b] AFMC (2010) report for medical students
[c] AFMC (2012) report for postgraduates

Embedded within this recommendation are three important motivations toward CBME (Hodges 2010):

- Increased efficiency
- Decreasing training time
- Reducing the overall cost of medical education

19.2.2.2 Defining and Identifying the Rationale for CBME

In this section, a collective definition of competency and its core elements are presented. The motivation for CBME and recommendations for its implementation are discussed.

Definition of competency. Although authors such as McGaghie et al. (1978) have argued that CBME cannot have a single definition, Frank et al. (2010a) conducted a systematic review of the literature to identify a unitary twenty-first century definition which is now in widespread usage by medical educators. Frank et al. defined CBME as "an approach to preparing physicians for practice that is fundamentally oriented to

Table 19.2 Rationale for implementing CBME and recommendations for implementation

Rationale for CBME	Recommendations for implementing CBME
1. Focus on curricular outcomes 2. Emphasis on abilities and competencies 3. De-emphasis on time-based training 4. Promotion of learner centeredness	1. Identify the abilities needed of graduates 2. Explicitly define the required competencies and their components 3. Define milestones along a development path for the competencies 4. Select educational activities, experiences, and instructional methods 5. Select assessment tools to measure progress along the milestones 6. Design an outcomes evaluation of the program

graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It deemphasizes time-based training and promises greater accountability flexibility, and learner-centeredness" (Frank et al. 2010a, p. 636). Supporting this definition, they identified eight components:

1. Defined outcomes and milestones (developmental levels that correspond to a competency framework)
2. Curriculum of competencies
3. Demonstrable/observable abilities
4. Assessment of competencies that indicate process toward defined outcomes
5. Learner-centered
6. Serving societal needs
7. Contrasting with time-based or process-based model
8. Implementation

Rationale for implementing CBME. Frank et al. (2010b) indicated that based on an international discussion and consensus, the rationale for CBME can be organized into four themes. To translate CBME into actual practice, the article also provided a six-step recommendation for planning CBME curricula. Table 19.2 summarizes these main points.

As scholars and practitioners in medical education consider the practical implications of adopting CBME into practice, an important challenge is the development of an assessment system to measure competencies and outcomes that have requisite quality and validity. The ensuing section discusses ideas that are emerging in the assessment of CBME and some of the challenges that lie ahead.

19.3 Assessments in CBME

To create a curriculum that focuses on outcomes and competencies and engages in learner centeredness, a multifaceted assessment system is needed (Holmboe et al. 2010). Furthermore, to create an assessment that adheres to the clinical work setting where medicine is practiced, WBAs need to be included in the assessment (Norcini and Burch 2007). The nature of CBME requires assessment to be continuous, frequent, criterion-based, and developmental, while using tools that generate validated inferences about and for learners. Within this context, Holmboe et al. (2010) outlined six components of an effective assessment system in CBME:

1. Assessments need to be continuous and frequent
2. Assessments must be criterion-based, using a developmental perspective
3. Competency-based medical education, with its emphasis on preparation for what the trainee will ultimately do, requires robust work-based assessment
4. Training programs must use assessment tools that meet minimum standards of quality
5. We must be willing to incorporate more "qualitative" approaches to assessment
6. Assessment needs to draw upon the wisdom of a group and to involve active engagement by the trainee

The authors also note that future assessments will need to delve into interactions between competence and clinical practice. Concerns raised include reducing CBME into "checkboxes" (Talbot 2004), assessing an overly large number of milestones, and the lack of appropriate assessment forms. Measuring team-based competence and associated outcomes has also been noted as a challenge that will need to be addressed, as the practice of medicine occurs in collaborative environments. At present, most measures of competence relate to individuals and measurable outcomes and appropriate assessments for teams that are not readily available (Hodges 2010). Finally, faculty development needs to be addressed, helping faculty to understand best practices for assessment in the era of CBME.

19.4 Challenges in CBME in the Face of Modern Psychometrics

As noted thus far in this chapter, CBME poses challenges to traditional notions of assessment and in particular the concepts of psychometrics that up to now formed the basis for assessment standards in medical education (AERA, APA, and NCME 2014). In a time-based curriculum, traditional point-in-time summative assessments can be analyzed using standard psychometric validity frameworks. However, in CBME, where time is no longer fixed, psychometric approaches to analyze data about competencies and outcomes are unclear, when they are gathered over time; to date, there is no general framework for measuring competence along a developmental (longitudinal) framework.

Competencies and entrustable professional activities (EPAs). Based on trends in CBME, new assessment methods will need to be developed to measure outcomes: (1) competencies and (2) EPAs. Assessments in CBME need to ensure standardized levels of proficiency that all graduates of the program have sufficiently attained before being deemed "competent" as defined by a competency framework.

The ACGME core competencies (medical knowledge [MK], patient care [PC], professionalism [PROF], interpersonal and communication skills [ICS], practice-based learning and improvement [PBLI], and system-based practice [SBP]) can serve as one such framework. Within this framework, "milestones" are fine-grained developmental levels associated with the competencies. For example, in Internal Medicine residency programs, "Gathers and synthesizes essential and accurate information to define each patient's clinical problem(s)" (PC1; see ACGME and ABIM 2014) is a milestone corresponding to Patient Care with five distinct developmental levels, ranging from "Critical Deficiencies" (Level 1) to "Ready for Unsupervised Practice" (Level 4), and "Aspirational" (Level 5). In CBME, the goal is for learners to achieve a minimum of Level 4 prior to completing the residency program.

EPAs indicate what the learners can actually do (as opposed to qualities represented in competencies) and are units of professional practice. In this sense, EPAs describe specific work that professionals are entrusted to perform as outcomes of their curricular experience. Some EPAs can require multiple competencies for executing the work. For example, "performing an appendectomy" will require a learner's competence in MK and PC; "executing a patient handover" will require competence in MK, PC, ICS, and SBP (ten Cate 2013).

As is evident from these examples, measuring competencies and EPAs in CBME will require using different assessment methods and combining scores from different assessments. Measuring milestones in PC or EPAs that require PC could involve aggregating information from various assessments, such as written tests, rotation evaluations, and even objective structured clinical examinations (OSCEs). The reason for combining scores from different assessment is that each assessment only provides a partial picture of the learner's performance; a full picture of their competence or entrustment can only be gained by aggregating their performance across a "system" of assessments targeted to measure the learner's ability (Holmboe and Hawkins 2008). Each assessment can have different characteristics including but not limited to psychometric validity evidence. Finding methods to optimally combine these scores in a "system" of assessment, one that measures global concepts of competence and includes entrustment is a challenge that must be resolved urgently given that, in the United States currently, Clinical Competency Committees (CCCs) are tasked with aggregating learner performance to determine overall competence.

Composite reliability and validity. One potential approach to aggregate assessment scores is to use a *composite validity* approach, in which information from multiple assessment tools is used to generate inferences about competence (Park et al. 2016). Psychometricians have for decades studied measurement characteristics of combining constructs from multiple assessments (i.e., assessment system) to create validity evidence, including composite score reliability (e.g., see Kane and Case 2004; Borsboom 2012; Markus and Borsboom 2013; de la Torre and Douglas 2004; Rao and Sinharay 2007). For example, if (1) rotation evaluations, (2) written tests, and (3) OSCEs were used to measure a milestone in PC or an EPA associated with PC, prior literature on the psychometric characteristics of the different

Table 19.3 Assessments used to measure Internal Medicine milestone PC1: example of composite score reliability

Assessment used to measure PC1: "Gathers and synthesizes essential and accurate information to define each patient's clinical problem(s)"	Reliability	Weight (%)	Composite reliability
Rotation evaluations	0.45	50	**0.72**
Written examinations	0.65	25	
OSCE	0.55	25	

Note Reliability and weights based on hypothesized psychometric characteristics and importance judged by the institution. Calculation of composite reliability is based on Kane and Case (2004), assuming correlations between assessments range between 0.35 and 0.50

assessment methods indicates that perhaps only written tests would have sufficiently high reliability, while rotation evaluations or OSCEs may have lower reliability when administered in local medical schools or residency programs (e.g., Park et al. 2014; Yudkowsky et al. 2014). By applying a composite score reliability approach, the composite psychometric characteristics of the overall aggregated PC score may be better than the reliability of the individual assessments. Table 19.3 demonstrates an example for the Internal Medicine milestone PC1 ("Gathers and synthesizes essential and accurate information to define each patient's clinical problem(s)").

As demonstrated in Table 19.3, the individual reliability indices of the three assessments vary between 0.45 and 0.65, which may not satisfy sufficient reliability levels for making decisions on learners (Nunnally 1978). However, when the three assessments are viewed as indicators that form a composite, the composite reliability will increase to 0.72. This idea of composite reliability extends to composite validity, where validity evidence for the composite assessment is maximized by using information from different scores.

While tools such as composite reliability and validity can be translated to CBME, it is worth noting that these methods were not originally developed in the health professions setting. Historically, the existing validity frameworks used in medical education (e.g., Kane or Messick; see Brennan 2006; Kane 2013) were motivated from general education contexts where multiple-choice tests are of prevalent use (and performance-based assessments are beginning to gain further use and appreciation; see Gordon 2008). These contexts differ significantly from WBAs that are a required part of CBME. As such, psychometric research and frameworks for CBME need to be developed or refined. In this regard, the importance of developing new psychometric frameworks should be considered a priority, rather than as a long-term project or left as an initiative in which only testing organizations invest.

Subjective, holistic, and narrative forms of assessment. An alternative approach to the recombination of quantitative test scores and psychometric analysis is the use of expert raters to integrate complex information to make decisions about learner competence. Gingerich (2015) has pointed out that that "entrustment" is a social judgment and as such, may be more amenable to integrated, holistic thinking on the part of supervisors than that are areas of knowledge and skills, domains which are more easily broken down into component scores. Similarly, van der

Vleuten et al. (2010), Kuper et al. (2007), Rawlings et al. (2015), and Whitehead et al. (2015) have argued that much more use can be made of qualitative paradigms and data in creating assessment systems appropriate to integrated notions of competence. Hodges (2013) has argued that it may be possible to use multisource, subjective assessments of competence to create composite profiles of competence that are both richer and more informative for further learning than simply combining a set of quantitative scores from diverse instruments.

19.5 Conclusion

This chapter provides a broad overview of CBME starting with the reforms rooted in Flexner's 1910 report. Origins, concepts, and limitations of time-based medical education curriculum, which began from university-based traditions of fixed time courses, are presented. Time-based medical education and its enduring appeal is presented from a historical perspective, noting that the development of competency frameworks such as the ACGME core competencies and of conceptual models such as Miller's Pyramid has motivated the need to replace time-based models with competency-based models.

Among key limitations of time-based models is the lack of flexibility to measure meaningful, clinical practice outcomes due to the fixed time structure that forces assessment to be standardized, point-in-time and generally at the end of rotations. Moreover, curricular reform has been noted to be difficult in traditional time-based programs. Recognizing these limitations, some educators have suggested that a focus on curricular outcomes, emphasis on abilities and competencies, de-emphasis on time-based training, and promotion of learner centeredness are reasons for moving toward CBME.

Practical recommendations have been identified for transitioning from theory to implementation of CBME in medical school curricula. Some of the guidelines for assessments in CBME have been presented here. We have given particular emphasis to challenges and unresolved issues related to assessment and CBME. Psychometric issues arising from assessments in CBME are among important challenges associated with measuring outcomes when curricular time is no longer fixed. Furthermore, some competencies and EPAs require combining scores from different types of assessments. To this end, the use of composite reliability and validity is proposed and discussed as is the possibility of holistic, subjective judgment models. However, we argue, composite reliability and validity and judgment models need to be better conceptualized and refined by the medical education community before being widely implemented. We recommend further research to advance current understanding of assessment systems within a CBME framework.

Overall, CBME, if widely implemented, will be an important and meaningful change to the way health professionals are trained and assessed. It may overcome some of the limitations of the time-based model, and may better serve public accountability, increase efficiency, and perhaps reduce unnecessary costs. As such, it may be part of a promising future in medical education. However, how well the competencies and outcomes—whether they are milestones or EPAs—can be assessed remains a challenge. Medical educators, psychometricians and those with expertise in judgment models need to develop new and innovative frameworks for addressing these challenges in measurement, as the CBME movement continues to accelerate.

Issues/Questions for Reflection

- Assessing competencies and EPAs in CBME will need continuous refinement to overcome challenges in meeting standards for validity
- Measuring team-based competence and associated outcomes will need to be examined and addressed in future assessments in CBME
- Psychometric concepts will need to be translated to meet CBME contexts, where assessments are longer administered in traditional point-in-time settings. Emerging solutions from narrative assessments and subjective judgment models may offer new insights
- Methods to allocate resources and support faculty to understand best practices for assessment will continue to require investigation

References

Accreditation Council for Graduate Medical Education & American Board of Internal Medicine. (2014). *The internal medicine milestone project*. Chicago, IL: ACGME.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards in educational and psychological testing*. Washington, DC: American Educational Research Association.

Association of Faculties of Medicine of Canada. (2010). *The future of medical education in Canada (FMEC): A collective vision for MD education*. Ottawa, ON: The Association of Faculties of Medicine of Canada.

Association of Faculties of Medicine of Canada. (2012). A collective vision for postgraduate medical education in Canada. Ottawa, ON: The Association of Faculties of Medicine of Canada.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research & Perspective*, *10*, 38–41.

Brennan, R. L. (2006). *Educational measurement* (4th ed.). Washington, DC: American Council on Education.

Cooke, M., Irby, D. M., & O'Brien, B. C. (2010). *Educating physicians: A call for reform of medical school and residency*. Stanford, CA: Jossey-Bass.

Crosson, F. J., Leu, J., Roemer, B. M., & Ross, M. N. (2011). Gaps in residency training should be addressed to better prepare doctors for a twenty-first-century delivery system. *Health Affairs, 30*, 2142–2148.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

Englander, R., Cameron, T., Addams, A., Bull, J., & Jacobs, J. (2015a). Understanding competency-based medical education. *Academic Medicine Rounds*. Retrieved from http://academicmedicineblog.org/understanding-competency-based-medical-education/

Englander, R., Cameron, T., Addams, A., Bull, J., & Jacobs, J. (2015b). Developing a framework for competency assessment: Entrustable professional activities (EPAs). *Academic Medicine Rounds*. Retrieved from http://academicmedicineblog.org/developing-a-framework-for-competency-assessment-entrustable-professional-activities-epas/

Flexner, A. (1910). *Medical education in the United States and Canada. A report to the Carnegie Foundation for the Advancement of Teaching* (Bulletin No. 4). Boston, MA: Updyke.

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., de Rossi, S., & Horsley, T. (2010a). Toward a definition of competency-based education in medicine: A systematic review of published definitions. *Medical Teacher, 32*, 631–637.

Frank, J. R., Snell, L. S., ten Cate, O., Holmboe, E. S., Carraccio, C., Swing, S. R., et al. (2010b). Competency-based medical education: Theory to practice. *Medical Teacher, 32*, 638–645.

Gidney, R. D., & Millar, W. P. J. (1994). *Professional gentlemen: The professions in nineteenth-century Ontario (Ontario Historical Studies Series)*. Toronto: University of Toronto Press.

Gingerich, A. (2015). What if the 'trust' in entrustable were a social judgment? *Medical Education, 49*, 750–752.

Gordon, E. W. (2008). The transformation of key beliefs that have guided a century of assessment. In C. A. Dwyer (Ed.), *The future of assessment* (pp. 3–6). New York, NY: Taylor & Francis Group LLC.

Gruppen, L. D., Mangrulkar, R. S., & Kolars, J. C. (2012). The promise of competency-based education in the health professions for improving global health. *Human Resources for Health, 10*, 1–7.

Hodges, B. D. (2010). A tea-steeping or i-Doc model for medical education? *Academic Medicine, 85*, S34–S44.

Hodges, B. D. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568.

Holmboe, E. S., & Hawkins, R. E. (2008). *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby Elsevier.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*, 676–682.

Institute of Medicine. (2014). *Graduate medical education that meets the nation's health needs*. Washington, DC: The National Academies Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kane, M. T., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240.

Kuper, A., Reeves, S., Albert, M., & Hodges, B. D. (2007). Assessment: Do we need to broaden our methodological horizons? *Medical Education, 41*(12), 1121–1123.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

McGaghie, W. C., Miller, G. E., Sajid, A. W., & Telder, T. V. (1978). *Competency-based curriculum development in medical education*. Switzerland: World Health Organization.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, S63–S67.

Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system. *New England Journal of Medicine, 366*, 1051–1056.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE guide no. 31. *Medical Teacher, 29*, 855–871.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Park, Y. S., Riddle, J., & Tekian, A. (2014). Validity evidence of resident competency ratings and the identification of problem residents. *Medical Education, 48*, 614–622.

Park, Y. S., Zar, F., Norcini, J., & Tekian, A. (2016). Competency evaluations in the Next Accreditation System: Contributing to guidelines and implications. *Teaching and Learning in Medicine*.

Rao, C. R., & Sinharay, S. (2007). *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, Netherlands: Elsevier.

Rawlings, A., Knox, A., Park, Y. S., Reddy, S., Williams, S. R., & Issa, N. (2015). Development and evaluation of standardized narrative cases depicting the general surgery professionalism milestones. *Academic Medicine, 90*, 1109–1115.

Starr, P. (1982). *The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry*. New York: Basic Books.

Talbot, M. (2004). Monkey see, monkey do: A critique of the competency model in graduate medical education. *Medical Education, 38*, 587–592.

Tekian, A., Hodges, B. D., Roberts, T. E., Schuwirth, L., & Norcini, J. (2015). Assessing competencies using milestones along the way. *Medical Teacher, 37*(4), 399–402.

ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of Graduate Medical Education*, *5*(1), 157–158.

ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Academic Medicine, 82*, 542–547.

van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. D. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology, 24*(6), 703–719.

Whitehead, C. (2010). Recipes for medical education reform: Will different ingredients create better doctors? *Social Science and Medicine, 70*, 1672–1676.

Whitehead, C. R., Kuper, A., Hodges, B., & Ellaway, R. (2015). Conceptual and practical challenges in the assessment of physician competencies. *Medical Teacher, 37*(3), 245–251.

Yudkowsky, R., Park, Y. S., Riddle, J., Palladino, C., & Bordage, G. (2014). Limiting checklist items to clinically-discriminating items: Improved validity of test scores. *Academic Medicine, 89*(7), 1–6.

Chapter 20
Assessment Challenges in Creating the Uniform Bar Examination: Politics, Practicality, and Psychometrics

Douglas R. Ripkey and Joanne Kane

Abstract In this chapter, we present an overview of the development and implementation of the Uniform Bar Examination (UBE). The UBE represents a shift in legal licensure policy toward standardization of the credentialing test across jurisdictions with the goal of benefitting students (through increased score portability), law schools (through increased standardization in examination materials), and jurisdictional boards of bar admissions (through increased efficiency). We begin with an overview of the practice and mission of the National Conference of Bar Examiners (NCBE)—the nonprofit organization that develops, creates, and oversees the UBE. In the next section, we review major services NCBE offers to jurisdictions and provide introductory information for readers unfamiliar with NCBE. The third section describes the general bar examination process, highlighting some challenges. In the final section, we describe the UBE in greater detail.

Takeaways

- Legal education in the United States has become increasingly standardized through ABA accreditation processes and state credentialing processes.
- Despite the strong tradition of local control over the bar admission process, there has been a notable movement toward convergence among the jurisdictions.
- The Uniform Bar Examination (UBE) represents a shift in legal licensure policy toward standardization of the credentialing test across jurisdictions with the goal of benefitting students, law schools, and jurisdictional boards of bar admissions.

D.R. Ripkey · J. Kane (✉)
National Conference of Bar Examiners, Madison, WI, USA
e-mail: jkane@ncbex.org

D.R. Ripkey
e-mail: dripkey@ncbex.org

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_20

- The purpose of the overall bar examination is to assess competence in matters that are considered essential knowledge and skills for the entry-level lawyer.

20.1 Background

The legal community formally recognized in 1921 that legal licensure candidates "… should be subject to an examination by public authority to determine [their] fitness" (American Bar Association 2015). Historically, entry into the profession has been focused locally with no central organization taking a direct leadership position in assessment of readiness to begin professional legal practice (Eckler 1996). "Fitness," "readiness," and "competency" were determined locally. Today, "competency" and "fitness" are more typically two separate hurdles in legal licensure, as we will see in this chapter. Each jurisdiction (i.e., state, territory, or District of Columbia) has made the determination of an applicant's competency to practice law in that jurisdiction based on the jurisdiction's own standards. The responsibility lies with the jurisdiction's Board of Bar Examiners. Jurisdictional Boards of Bar Examiners can be politically independent, under the wing of the jurisdiction's supreme courts, and/or connected to the local bar associations. Currently, there are 56 separate jurisdictions that have control over admission to their local bar. Although each of the 56 separate jurisdictions has its own single bar examination agency, these agencies vary widely in terms of staffing level, monetary resources, and testing/psychometric expertise in conducting licensure functions.

For most of its history, the private, nonprofit National Conference of Bar Examiners (NCBE) has served in the roles of educational facilitator and change agent within the assessment realm of the legal profession. At the time of its formation in 1931, NCBE was charged with helping state boards of bar examiners cooperate with each other, with the law school community, and with the organized bar (Karger 1996). At that time, the concept of a written bar examination administered by a central board of bar examiners within a statewide jurisdiction was a novel idea resulting from the then-new practice of preparing for a legal career by studying a prescribed curriculum within a law school rather than the time-honored model of following an individual self-study program with extensive mentoring.

The goal in founding NCBE was a facilitative one; it was hoped that bar examiners would share information and experiences, and through this process begin to adopt shared best practices. In fact, part of NCBE's formal charge was

> increasing the efficiency of the state boards in admitting to the bar only those candidates who are fully equipped both from a standpoint of knowledge and of character to serve as lawyers, and also to study and to cooperate with the other branches of the bar in dealing with problems of legal education (Karger 1996, p. 14).

At present, the self-defined mission of NCBE is

> to work with other institutions to develop, maintain, and apply reasonable and uniform standards of education and character for eligibility for admission to the practice of law and to assist bar admission authorities by: 1) providing standardized examinations of uniform and high quality for the testing of applicants for admission to the practice of law, 2) disseminating relevant information concerning admission standards and practices, 3) conducting educational programs for the members and staffs of such authorities, and 4) providing other services such as character and fitness investigations and research" (National Conference of Bar Examiners 2015).

Since the founding of NCBE nearly a century ago, legal education in the United States has become increasingly standardized through ABA accreditation processes and state credentialing processes, including the now widespread adoption of the Multistate Bar Examination and increasingly widespread adoption of the Uniform Bar Examination (UBE).

20.2 NCBE Services

NCBE offers a variety of targeted services to bar admission authorities. The sections below provide an overview of these services.

20.2.1 Character and Fitness

NCBE conducts character and fitness investigations on applicants seeking either admission to the bar or a limited license to practice law. These investigations are completed at the request of participating jurisdictions. Although some jurisdictions have a specialized application, most require candidates to complete a 28-question online common application that collects background information including personal references, education records, employment records, interactions with the criminal justice system, records of noncriminal sanctions and complaints, and other information regarding fitness to serve as a capable and ethical professional. NCBE staff members verify the information and produce a standardized candidate report for the requesting jurisdiction. Currently, more than half of all jurisdictions participate directly in this service and nearly 75 % participate in an ancillary service that allows candidates' history of bar examination participation to be cataloged in a central NCBE database. Recently, NCBE has also worked with jurisdictions to develop a model instrument to be used for evaluation of Americans with Disabilities Act (ADA) accommodation requests.

20.2.2 Educational Conferences

A major activity of NCBE is its direct sponsorship and organization of three yearly educational seminars that are held in spring and fall in rotating locations across the United States. Using NCBE affiliated experts and other nationally renowned presenters, the seminars are designed to provide continuing education and guidance for bar examining and admission authorities from all jurisdictions. For example, at the April 2015 seminar in Chicago, more than 381 individuals representing 54 of 56 jurisdictions had the opportunity to hear speakers in 19 different sessions discuss topics including character and fitness issues, the latest in cheating detection, the impact of the UBE, maintaining grading consistency, and admitting foreign-trained lawyers. NCBE also provides other periodic seminars to smaller groups on more specialized topics.

20.2.3 Publications and Communications

NCBE sponsors the development and publication of materials that foster the sharing of information among constituencies. For example, NCBE hosts a listserv specifically designed to allow bar examination administrators to interact and exchange information. The listserv also allows NCBE to communicate directly in response to questions or issues that arise within a jurisdiction or across jurisdictions.

The Supervisor's Manual is a detailed document that outlines best practices for handling the logistical aspects of test administration. Developed over the course of many years with the input of NCBE staff and experienced bar administrators, it covers pre-administration preparations (receipt and storage of materials, seating practices, registration), activities on the test day (material distribution, standard conditions, proctor responsibilities, examinee instructions), and post-administration tasks (security, shipping, incident reporting).

The Comprehensive Guide to Bar Admissions, which has been published annually since the early 1980s, is a joint effort of NCBE and the American Bar Association (ABA) Section of Legal Education and Bar Admissions. Targeted broadly to both individuals and institutions, this publication (available in paper and electronic formats) is designed to be a comprehensive source for descriptive and comparative information about admission requirements in each of the 56 jurisdictions. It includes eligibility requirements for new and established lawyers, bar examination structure (e.g., timing, components, and scoring), rules for entry in lieu of examination, and requirements for continued practice. Information is provided by individual jurisdictions' bar examination administrators and is compiled and edited by NCBE staff.

The Bar Examiner is a quarterly publication which has been in print since 1931. With a circulation of about 3400, it is published as a service to members of the courts (judges and state Supreme Court members), legal academia (law school faculty members, Deans, and libraries), bar admission administrators, members of bar examining boards and character committees, and others with special interest in the bar admission process. In addition to ongoing standard features like the NCBE President's and Chair's pages, the Testing Column, and Litigation Updates, each issue has a collection of articles covering a wide range of topics which have included evaluations of character and fitness, ADA accommodations, testing practices (including development, administration, grading and scoring), diversity issues, globalization and foreign legal education, the bar examination and its alternatives, plus trends and statistics. At a very practical level, this publication allows NCBE to communicate with and educate stakeholders on key issues by articulating best practices and reporting research findings.

20.2.4 Research

Research has played an increasingly critical role in the fulfillment of NCBE's mission. NCBE staff members engage with individual jurisdictions as well as groups of jurisdictions to investigate topics of mutual interest. Often this collaboration has been in response to a specific request from a jurisdiction with an issue they wanted to investigate or resolve (Case 2007a). For example, it has been common for NCBE to conduct studies evaluating the impact of potential changes to jurisdictions' scoring models, assessing the performance of bar examination graders, and/or analyzing performance differences among various groups. These results provide information and insights that are useful to the jurisdiction, NCBE, and the larger bar examination community. Although the detailed research reports are typically considered confidential, the individual jurisdiction benefits by getting their question/concern evaluated and NCBE has an empirical basis to make a general best practice recommendation which can be shared with other jurisdictions. As a complement to conducting research in response to inquiries from jurisdictions, NCBE has asked jurisdictions to participate in research projects that are geared toward both the identification of best practices and the evaluation of validity evidence for the bar examination itself. These investigations have included such topics as the relationship among scores on the bar examination components (Case 2008b; Ripkey and Case, April 2009), threats to validity posed by cheating (Albanese and Wollack 2013, October), performance comparisons among measures of academic success (Mroch and Ripkey, April 2007; Ripkey and Mroch 2007), and differential performance by candidate groups (Case 2006b; Kane et al. 2007; Ripkey, March 2008).

As more data has become available internally through the expansion of testing activities, NCBE has conducted operational-level research as well including evaluations regarding rounding (Case 2004a; Kane 2003), equating (Kane et al. 2010; Mroch et al. 2010; Suh et al. 2010), security analyses of gain scores (Case 2005b), security analyses of response pattern similarity (Lee and Albanese 2013, October), subscores (Albanese 2014), and repeater patterns (Case 2007b). In terms of volume, NCBE has, over the past decade, completed more than 50 studies providing direct support for jurisdictions.

In addition to operational research, NCBE supports research of general interest to measurement professionals. One major focus of these externally relevant projects has been cheating detection and disposition (Albanese 2013, April; Albanese et al. 2013, October; Albanese and Wollack 2013, October; Lee et al. 2014, October; Tracy and Albanese 2014, October). Another recent broad research focus has been the detection of rater bias in high-stakes licensing examinations (Albanese and Ripkey 2012, April). As NCBE's testing and research departments expand, so too does the capacity to produce increasingly sophisticated research; this research is used for operational purposes, jurisdictional purposes, and broader scientific purposes.

20.2.5 Expanding Role

While NCBE has continued to act in its historical roles as supporter and educator for the disparate licensure groups through their fitness evaluations, sponsorship of educational meetings, production of periodicals, and facilitating research, its horizons have also expanded over time. NCBE has been on a trajectory of shifting from a supporting role in bar examinations into becoming a major contributor as it has gradually evolved into a fully fledged testing organization. At first, outside test vendors like ETS and ACT provided most of the test development and scoring services, with NCBE in a general oversight role. NCBE first began to expand their test development presence by moving into the drafting and test construction functions. In subsequent years, NCBE expanded into the psychometric side by hiring in-house psychometric staff to conduct the equating calculations for some of the bar examination components it developed. Currently, all of the equating, scoring, and score reporting functions for NCBE's multiple-choice exams are conducted internally.

20.2.6 Bar Examination Materials

Today, NCBE develops four tests that are available to jurisdictions for their discretionary use as part of the local bar admission process.

The Multistate Bar Examination **(MBE)** assesses the extent to which an examinee can apply fundamental legal principles and legal reasoning to analyze given fact patterns. The MBE currently contains 200 multiple-choice questions, 190 of which are scored; the remaining 10 are used for purposes of pretesting. The 190 scored questions on the MBE are distributed among seven primary content areas: Civil Procedure (27), Constitutional Law (27), Contracts (28), Criminal Law and Procedure (27), Evidence (27), Real Property (27), and Torts (27). Each question requires examinees to choose the best answer from four alternatives. All questions are designed to be answered according to generally accepted fundamental legal principles, unless noted otherwise in the question (National Conference of Bar Examiners 2011a).

MBE questions are developed by drafting committees composed of recognized experts in the various subject areas. Each MBE item undergoes a multistage evaluation process over the course of several years. Besides intensive review by the drafting committee members and testing specialists, each question is assessed by other national and state experts. All test questions must successfully pass all reviews before they are included in the MBE. After an MBE is administered, and before scores are released, the performance of each test question is examined by content and testing experts. This final review is conducted to ensure that the exam is graded fairly, particularly with regard to any questions affected by recent changes in the law (Hill 2015).

The MBE is administered by participating jurisdictions on the last Wednesday in February and July of each year. On each administration, the MBE is divided into morning and afternoon testing sessions of 3 hours each, with 100 questions in each session. MBE answer sheets are scanned and centrally scored. A total equated score (on a 200-point scale) is calculated using a three-parameter IRT model (Kane and Mroch 2005). Each jurisdiction determines its own policy with regard to the weight given to the MBE in relation to any other examination components in determining passing outcome on their local bar examination.

The MBE was first administered in 1972 by only 19 jurisdictions. The MBE is now used by 49 states (Louisiana does not administer the MBE), plus the District of Columbia, Guam, the Northern Mariana Islands, Palau, and the U.S. Virgin Islands. Despite the strong reliability and increasingly widespread use of the MBE, some individuals and jurisdictions have had misgivings about the value of multiple-choice questions. To address some of these concerns and in recognition of the fact that some knowledge, skills, and abilities required of newly licensed attorneys might be more appropriately measured by other test formats, NCBE developed two additional tests: the Multistate Essay Examination and the Multistate Performance Test (Gundersen 2007).

The Multistate Essay Examination **(MEE)** is designed to test an examinee's ability to (1) identify legal issues raised by a hypothetical factual situation; (2) separate material which is relevant from that which is not; (3) present a reasoned analysis of the relevant issues in a clear, concise, and well-organized composition; and (4) demonstrate an understanding of the fundamental legal principles relevant to the probable solution of the issues raised by the factual situation. The primary distinction between the MEE and the Multistate Bar Examination (MBE) is that the MEE requires the candidate to demonstrate an ability to communicate effectively in writing (National Conference of Bar Examiners 2011b).

More than 100 people are involved in the development of the MEE. Members of the MEE Drafting Committee and other outside experts draft original questions for the MEE. Academics and practitioners, who are experts in the fields covered by the test, serve as drafters. Questions are edited by the Drafting Committee, pretested, analyzed independently by outside experts, and reviewed by the boards of bar examiners in the jurisdictions that use the test and members of the MEE Policy Committee. Finally, the test is revised by the Drafting Committee.

NCBE offers six 30-minute questions per administration cycle. Questions on a given test form may be drawn from the MBE subject areas (Civil Procedure, Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts) and/or from other subject areas (Business Associations, Conflict of Laws, Family Law, Trusts and Estates, and the Uniform Commercial Code). Some questions may require analysis of more than one subject area (Gundersen 2006). Depending upon jurisdictional practices and resources, responses are recorded and submitted electronically (on laptops) or on paper.

MEE questions are graded locally. Graders use extensive analyses prepared by NCBE to assist them. Each analysis identifies the issues raised by the question, cites appropriate authority, and indicates suggested weights for crediting discussion of the issues. NCBE sponsors a grading workshop on the weekend following test administration. Separate grading sessions are held for each question and led by drafting committee members who are seasoned question writers. Graders for each question meet to review the analysis and sample answers to each question. Revisions to the analyses are sent to each user jurisdiction immediately after the grading workshop. Copies of past MEEs are made available publicly on the NCBE website (www.ncbex.org).

Although NCBE has a recommended score scale, a 1–6 holistic evaluation, user jurisdictions are free to define their own grading methodology and score scale. In addition, each jurisdiction determines its own policy with regard to the relative weight given to the MEE and any other scores in determining passing status on the local bar examination.

***The Multistate Performance Test* (MPT)** is developed as a performance test rather than a test of substantive knowledge. It is designed to test an examinee's ability to use fundamental lawyering skills in a realistic situation. Each prompt elicits a response used to evaluate an examinee's ability to complete a task that a beginning lawyer should be able to accomplish. The MPT requires examinees to (1) sort detailed factual materials and separate relevant from irrelevant facts; (2) analyze statutory, case, and administrative materials for applicable principles of law; (3) apply the relevant law to the relevant facts in a manner likely to resolve a client's problem; (4) identify and resolve ethical dilemmas, when present; (5) communicate effectively in writing; and (6) complete a lawyering task within time constraints. These skills are tested by requiring examinees to perform one or more of a variety of lawyering tasks. For example, examinees might be instructed to complete any of the following: a memorandum to a supervising attorney, a letter to a client, a persuasive memorandum or brief, a statement of facts, a contract provision, a will, a counseling plan, a proposal for settlement or agreement, a discovery plan, a witness examination plan, or a closing argument (Gundersen 2007; National Conference of Bar Examiners 2011d).

The MPT is developed by the MPT Drafting Committee, which has extensive experience in writing, editing, and grading performance test items. All MPTs are pretested, critiqued by independent experts, and reviewed by the boards of jurisdictions using the test prior to final revision by the Drafting Committee. MPT drafters are legal clinicians, practitioners, or judges.

NCBE offers two 90-minute MPT items per administration cycle. A jurisdiction may select one or both items to include as part of its bar examination. The materials for each MPT include a file and a library. The file consists of source documents containing all the facts of the case. The specific assignment the examinee is to complete is described in a memorandum from a supervising attorney. The file might also include transcripts of interviews, depositions, hearings or trials, pleadings, correspondence, client documents, contracts, newspaper articles, medical records, police reports, or lawyer's notes. Relevant as well as irrelevant facts are included. Facts are sometimes ambiguous, incomplete, or even conflicting. As in practice, the records or a client's version of events may be incomplete or unreliable. Examinees are expected to recognize when facts are inconsistent or missing and are expected to identify sources of additional facts. The library may contain cases, statutes, regulations, or rules, some of which may not be relevant to the assigned lawyering task. The examinee is expected to extract from the library the legal principles necessary to analyze the problem and perform the task. Because the MPT is not a test of substantive law, the library materials provide sufficient substantive information to complete the task. Depending upon the jurisdictional practices and resources, responses are recorded and submitted either electronically (on laptop) or on paper (Bosse 2011).

Each MPT is accompanied by grading guidelines designed to assist jurisdictions in scoring the test. NCBE sponsors a grading workshop on the weekend following test administration in conjunction with the MEE Grading Workshop. Similar to the process for MEE scoring, jurisdictions score the MPT. Jurisdiction-specific policy determines what weight to give the MPT in relation to the other parts of their bar examinations. Copies of past MPTs and their associated scoring guidelines are available on the NCBE website (www.ncbex.org).

The Multistate Professional Responsibility Examination **(MPRE)** measures examinees' knowledge and understanding of established standards related to a lawyer's professional conduct; the MPRE is not a test to determine individuals' personal ethical values (National Conference of Bar Examiners 2011c). The MPRE is administered separately from the bar examination, and is required by most jurisdictions for admission to the practice of law. Lawyers serve in many capacities: for example, as judges, advocates, counselors, and in other roles. The law governing the conduct of lawyers in these roles is applied in disciplinary and bar admission procedures; by courts in dealing with issues of appearance, representation, privilege, disqualification, and contempt or other censure; and in lawsuits seeking to establish liability for malpractice and other civil or criminal wrongs committed by a lawyer while acting in a professional capacity. The MPRE is based on the law governing the conduct of lawyers, including the disciplinary rules of professional conduct currently articulated in the ABA Model Rules of Professional Conduct (MRPC) and their Model Code of Judicial Conduct (CJC), as well as controlling constitutional decisions and generally accepted principles established in leading federal and state cases and in procedural and evidentiary rules (American Bar Association Center for Professional Responsibility 2013; American Bar Association 2010).

The MPRE is developed by a six-member drafting committee composed of recognized experts in the area of professional responsibility. Before a test question is selected for inclusion in the MPRE, it undergoes a multistage review process that occurs over the course of several years. Besides intensive review by the drafting committee and testing specialists, each test question is reviewed by other experts. All test questions must successfully pass all reviews before they are included in the MPRE. After an MPRE is administered, the statistical performance of each test question is reviewed and evaluated by content and testing experts before the questions are included in the computation of examinees' scores. This final statistical review is conducted to ensure that each test question is psychometrically sound (Case 2004b).

The MPRE is offered three times per year at established test centers across the country. It includes 60 multiple-choice questions—50 are scored and 10 are non-scored pretest questions. Each MPRE question provides a factual situation along with a specific question and four possible answer choices. The reported MPRE score is an equated standard score ranging from 50 to 150. Equating is conducted using a three-parameter IRT model. Minimum passing scores are established by each jurisdiction; an equated scaled score of 75 is the most common value.

20.3 General Bar Examination Process

20.3.1 Contents and Administration

The purpose of the overall bar examination is to assess competence in matters that are considered essential knowledge and skills for the entry-level lawyer. As a credentialing examination, the bar examination is targeted at minimum competence and is developed to protect both the public and the profession from poorly qualified practitioners.

The MPRE, which is typically taken by candidates while they are still in law school, is a stand-alone evaluation of professional ethics that is required for bar admission in 54 of 56 jurisdictions.

Almost all jurisdictions' bar examinations (i.e., 54 of 56) have two general elements: (1) the multiple-choice MBE and (2) the constructed response "written test" consisting of a mixture of MPTs, MEEs, and/or other locally developed essay questions. As outlined above, there is overlap among these examination components in terms of the knowledge, skills, and abilities they assess.

Most jurisdictions administer the MBE and some combination of the MEE, MPT, and local essays as part of a multiday test administration that is given each year at the end of February and July. Examinees generally sit for an administration during or after their third (i.e., final) year of law school. They must register with the jurisdiction in which they are seeking admission, and it is the jurisdiction that administers the test materials provided to them from NCBE and any other locally developed examination materials (e.g., local essays). NCBE provides MBE score results to jurisdictions who then evaluate examinee performance on this component in conjunction with the performance on the other components of the bar examination.

20.3.2 Scoring Models and Practices

Jurisdictions have complete autonomy in choosing a scoring model. However, because almost all jurisdictions use two components (i.e., MBE and Written), they follow either a compensatory or non-compensatory scoring rule. With a *non-compensatory* scoring rule, the separate component scores are considered independently, and it is necessary to meet the minimum passing standard on each component in order to achieve success on the examination. For a *compensatory* scoring rule, the scores on the separate components are combined into a single overall score, which is then used as the basis for a decision. A non-compensatory score rule tends to make sense in cases where the skills being measured are (1) clearly distinct with little overlap in what is being measured; (2) each necessary for effective performance; and (3) each measured with adequate reliability. If any one of these conditions fails, a compensatory scoring rule is recommended. Compensatory rules are often preferred because they are likely to be considerably more reliable than non-compensatory rules (Kane 2009).

NCBE recommends a compensatory scoring model. Conceptually, there is overlap among the examination components in terms of the knowledge, skills, and abilities they assess. Each of the components has a somewhat different purpose, but together they are designed to measure the extent to which examinees have the requisite knowledge and skills to be licensed to practice law. NCBE's public (Ripkey and Case, April 2009), operational, and private research shows scores on the two components are moderately correlated (typically attenuated correlations in the 0.60 s, but can range from the 0.40's to the 0.80's) and that reliability of the written component varies across jurisdictions (ranging from about 0.60's to 0.70's); this range is much lower than the 0.90 or higher reliability value typically seen on the MBE. Of the 53 jurisdictions that use both the MBE and written components, five use a non-compensatory score rule in which the total MBE and the total written component scores are evaluated separately using their locally defined minimum performance standard for each. The remaining 48 jurisdictions use a compensatory scoring model to form a total composite score (National Conference of Bar Examiners and American Bar Association 2015).

Jurisdictions using compensatory scoring must account for the differences in score scale use for the components, how much weight each component contributes to the composite score, and the consistency of score interpretation across time.

Since jurisdictions often use a 1–6 grading scale for each case within the written component, the total raw written score (i.e., the sum of the raw score on the individual written cases) is on a much different scale than the 200-point MBE total score. For the proper weighting to occur, total scores from the jurisdiction's two components must be placed on the same score scale.

To achieve this, NCBE's recommendation is that within a jurisdiction, raw total written scores should be converted by linear transformation into their MBE scale—a process known colloquially as "scaling the essays to the MBE" and/or the "standard deviation method" of scaling. Functionally, within jurisdiction and administration, the raw total essay scores are converted to a standardized score (i.e., z-score) and then linearly "scaled" by assigning a value equivalent to the same z-score location on the MBE distribution for that same set of examinees. The formula can be expressed generically as follows:

Scaled Total Essay Score = Scaled MBE Mean + ((Scaled MBE SD/Raw Written SD) × (Raw Written Score − Raw Written Mean)).

This procedure has the effect of indirectly equating the written scores over time based on MBE performance, which is equated. Despite the change in scale, the rank ordering of jurisdictions' examinees on the written component is not altered, but the actual (i.e., scaled) scores have incorporated the adjustment necessary to reflect jurisdictions' cross-administration fluctuations in the difficulty of the essay questions, the severity of the graders, and/or the general proficiency of the examinee population. Ultimately, the process of "scaling the essays to the MBE" ensures consistency in the application of performance standards across administrations by correcting for cross-administration differences while preserving both the basic meaning of the total written scores and the relative standing of examinees on their written scores (Case 2005a, 2006a).

Once the two component scores are on the same (equated) scale, they are combined using the appropriate weighting. Again jurisdictions are free to create their composite score using any weighting they choose. Collectively, NCBE research for jurisdictions has shown that weighting the MBE score at least 50 % optimizes the total score's psychometric properties (Case 2008a) because of its high reliability and relatively strong relationship with the written component score (see the Kane and Case model for creation of weighted composite scores [Kane and Case 2004]). Currently, 33 of 48 jurisdictions that use a compensatory rule apply equal weighting (i.e., equated MBE score 50 % and scaled Written Score 50 %) when combining scores. Those jurisdictions that do not use equal weighting tend to provide greater weight (i.e., 60–67 %) to the written component (National Conference of Bar Examiners and American Bar Association 2015). Roughly translated to the 200-point MBE scale, the standard for the total composite score ranges from 129 to 145; a value of 135 is most common (Case 2011).

20.3.3 Commonality in Assessment

Despite the strong tradition of local control over the bar admission process, there has been a notable movement in the past decade toward convergence among the jurisdictions. In close partnership with jurisdictions and the bar examination community, NCBE has been able to identify and disseminate a series of best practices regarding the full gamut of the testing process. Similarly, NCBE has engaged both internal and external experts in the process of developing a collection of high-quality test instruments. At the end of 2014, 54 jurisdictions used the MBE, 41 jurisdictions used the MPT, and 31 jurisdictions used the MEE. There were 29 jurisdictions that administered all three components. All who used the three components also had a compensatory scoring model with most placing 50 % of weight on the MBE component and 50 % on the collective written component (i.e., MPT, MEE, and/or local essays) (National Conference of Bar Examiners and American Bar Association 2015). Given the high stakes involved with legal licensure, the increased knowledge of testing practice, and the availability of high-quality test materials, jurisdictions' bar examination practices have gradually converged.

20.4 The Uniform Bar Examination

20.4.1 Reasons for the UBE

There are many benefits associated with having a common multimodal assessment designed to produce a single score that could be used for admission consideration across jurisdictions. Collectively, a UBE with a transportable score positively impacts law schools, their graduates, and boards of bar admission; however, these benefits differ by constituency (Case 2010d).

Law schools whose graduates ultimately seek admission across various jurisdictions currently have some challenge in preparing students for a wide number of variations on those bar examinations. The benefit of a UBE is that all students (across schools and jurisdictions) face exactly the same exam for licensure and receive scores that have the same meaning across the country. The need for customization in preparation of students is thereby greatly reduced.

Recent law school graduates seeking to expand their employment opportunities by gaining admission to more than one jurisdiction can face the challenge of taking the bar examination in multiple places and dates under the current system because bar admission is granted on a jurisdiction-specific basis. However, the UBE notion of the transportability of a score across jurisdictions provides a significant benefit to an examinee who fails to get the job he or she intended and has to move to another jurisdiction to find work, or one who ends up working for a firm that has clients in multiple jurisdictions.

Jurisdictions benefit by a conservation of resources based upon adoption of a centrally developed UBE using instruments that meet professional testing standards. Because money, time, and expert resources are limited, many bar admission offices are burdened by the necessity of developing written exams and grading materials and of completing the development of these materials in a timely manner. Given that few bar admission authorities have testing or psychometric professionals on staff, it has not been uncommon for criticism to be raised alleging that some of the materials were not well written (VandeWalle 2009). Under the UBE, jurisdictions are relieved of this burden while still having the ability to offer high-quality examination instruments.

20.4.2 Concerns About the UBE

Although many members of the legal education and bar examination communities recognized the benefits of a UBE, there were also some significant concerns. The primary theme behind these concerns related to losing both local control of the admission process and specialized, locally relevant, exam content (Pogers 2010, December; VandeWalle 2009).

20.4.3 Pathway to the UBE

Since at least 2002, various constituencies (including the organized bar, bar examiners, courts, and legal educators) have recognized the increased similarity across jurisdictions in bar examination practices and the practical realities of changing legal practices and opportunities. These constituencies began formally questioning whether a uniform bar exam and its expected pooling of resources would improve the reliability and validity of state bar exams and better meet the needs of law schools with their national student bases and law school graduates with their multijurisdictional practices.

In 2002, representatives from NCBE, the American Bar Association (ABA), the American Association of Law Schools (AALS), and the Conferences of Chief Justices (CCJ) formed the Joint Working Group on Legal Education and Bar Admission. The activities of the Joint Working Group's activities resulted in an ABA-sponsored initiative for exploring a UBE. In the same time period, NCBE's Long Range Planning Committee decided that NCBE had a role in evaluating the feasibility of a uniform bar exam. A special subcommittee was formed. It eventually concluded that serious consideration should be given to the development of a uniform bar exam: this exam would include the MBE, MEE, and MPT, and apply a common testing, grading, scoring, and combining protocol. Like the ABA subgroup, the NCBE Special Committee on the Uniform Bar Exam acknowledged, however, that the concept of uniformity needed to be defined in greater depth to address and resolve many lingering questions and concern about details related to retention of local control.

To further explore this proposition, the NCBE Special Committee sponsored a conference in January 2008 attended by representatives of 21 jurisdictions, including 10 Supreme Court Justices and 17 chairs and administrators from individual state examining boards. These "interested parties" were primarily from jurisdictions using the full mix of NCBE developed exams (MBE, MEE, and MPT). After very open discussions that examined some concrete concerns, the group generally favored the development of a uniform bar exam. The NCBE Special Committee presented a specific written proposal to the jurisdictions in January 2009 (Thiem 2009). After several iterative refinements by the Special Committee and NCBE staff, another set of geographically centered special meetings of self-identified potential user jurisdictions was convened later that year and into the next to present a final model and gain "buy-in."

20.4.4 The UBE Model

The UBE is an NCBE developed examination composed of the three elements (MBE, MEE, and MPT) already in common use by the jurisdictions. However, in some ways the UBE is more than just a shared set of test components. It also represents the portability of the candidate's performance scores on a standardized measure. This allows for simplification of cross-jurisdictional licensure of new lawyers. It is a functional agreement to recognize the validity of test scores generated in any participating jurisdiction, predicated upon the fact that all jurisdictions offering the UBE administer the common elements under standardized administration protocols, score the components in a consistent way, and form a composite score in exactly the same way (Case 2009). Before scores are certified as official UBE scores, all participants must agree to adhere to specific policies (Early 2011) that are based on NCBE's previously established set of best practices (Case 2005a, 2008a, 2010b, c).

20.4.4.1 Standard Administration

In terms of standardized administration, UBE jurisdictions are required to follow the instructions set out in the Supervisor's Manual for administering the examination. In terms of composition, the morning of the first day of the UBE administration consists of a common set of six MEE questions, each of which must be completed within 30 min and answered according to generally applicable principles of law rather than jurisdiction-specific law. In the afternoon, the UBE jurisdictions administer the two MPT tasks in one seamless 3-hour test session rather than the two 90-minute sessions in the traditional model. The second day of the UBE consists solely of the MBE.

To create accurate transcripts for applicants who take the UBE multiple times or in multiple jurisdictions, NCBE must have sufficient biographical data to tie all the scores together. Thus, UBE jurisdictions agree to instruct applicants to provide the necessary identifying information on their MBE answer sheets.

20.4.4.2 Standard Grading

The MEE and MPT are scored in essentially the same way as they are scored for the traditional bar examination. Although applicants' answers are graded within jurisdiction using the locally established raw grading scale, UBE graders must adhere to the guidelines set out in the grading materials so that the same weight is assigned by all UBE jurisdictions to the various issues tested by each question. UBE graders agree to use general principles of law (rather than jurisdiction-specific principles) as identified in the grading materials prepared by the NCBE committees. UBE jurisdictions are not required to have their graders attend NCBE's grading workshop (either in person or by teleconference) the weekend following the examination, but attendance is encouraged.

20.4.4.3 Standard Scoring

Using the same "standard deviation method" as the traditional bar examination, jurisdictions' raw written component scores (regardless of jurisdictional variation in MEE/MPT grading scales) are scaled to the MBE scores to preserve the rank ordering of candidate performance while placing all scores on the same scale (which is equated indirectly because of the MBE). To maintain scoring consistency and comparability of scores, all UBE jurisdictions will adhere to a compensatory scoring model in which the MBE scores and the combined MEE/MPT scores will be weighted equally. Based on NCBE research (Case 2008b; Ripkey and Case, April 2009), the time allocations, and existing jurisdictional practices (National Conference of Bar Examiners and American Bar Association 2015), the precise weightings are MBE-50 %, MEE-30 %, and MPT-20 % in forming the total composite score which is portable across UBE jurisdictions. To help ensure the successful implementation of the scaling and score combination algorithms, NCBE performs all UBE scaling calculations for jurisdictions.

To earn UBE scores, applicants must sit for all portions of the examination in the same administration and cannot rely upon a component score that is banked or transferred. (A "banked score" is a component score earned in a prior examination in the testing jurisdiction where the applicant did not pass the exam but scored high enough on one component so as not to have to retake that component. A "transferred score" is one earned in a prior examination in another jurisdiction, where the applicant may or may not have passed depending on the requirements set by the receiving jurisdiction for accepting transferred scores.)

20.4.4.4 Local Practices and Autonomy

Essentially, it is the UBE score that is transportable—not bar admission status. All policies related to the requirements for admission on the basis of a transferred UBE score are left to the jurisdictions to be set independently. That is, the UBE score represents a demonstrated level of performance that can be trusted to have a consistent meaning across test administrations and locations, but the UBE performance standard required for admission remains completely under local control. Even if a candidate has a UBE score that meets a jurisdiction's minimum passing standard, licensure is not assured. The receiving jurisdictions may have a number of other conditions (e.g., educational background, a specific MPRE score, demonstrated character) which must be met prior to admission in that location.

By definition, the UBE does not include specific local content that is of unique interest in a particular jurisdiction. Jurisdictions that choose to use the UBE could assess applicant knowledge of local law in at least three ways: (1) requiring applicants to take and pass a course (similar to a CLE effort); (2) requiring applicants to pass a separate test on local content that could be administered at any time and would be scored separately from the UBE and treated as a separate hurdle; or (3) requiring applicants to take a separate test on local content just before or after the UBE is administered, which would be scaled to the UBE and could be combined with the UBE score for the purposes of making the local admissions decisions (Case 2010a). No jurisdictions are currently opting for the third alternative.

20.4.5 UBE Launch

At its annual meeting in July 2010, the Conference of Chief Justices adopted a UBE endorsement resolution proposed by its Professionalism and Competence of the Bar Committee (National Conference of Bar Examiners 2012). A similar endorsement was soon after approved by ABA Council of the Section of Legal Education and Admissions to the Bar in August 2010 (National Conference of Bar Examiners 2012). Both resolutions urged the bar admission authorities in each state and territory to consider participating in the development and implementation of a UBE.

In February 2011, Missouri and North Dakota were the first two states to administer the UBE and accept UBE scores; Alabama followed soon after—starting in July of the same year. As of 2015, 14 jurisdictions administer and accept UBE scores (Alabama, Alaska, Arizona, Colorado, Idaho, Minnesota, Missouri, Montana, Nevada, New Hampshire, North Dakota, Utah, Washington, and Wyoming) (National Conference of Bar Examiners and American Bar Association 2015). Kansas will administer its first UBE in February 2016 (National Conference of Bar Examiners 2015) and New York in July 2016 (The New York State Board of Law Examiners 2015). Iowa and Vermont have also recently announced that they will become UBE jurisdictions in 2016. As of July 2016, 18 of 54 jurisdictions will administer the UBE.

As the UBE matures, NCBE will continue to work with any jurisdiction that contemplates potential introduction of the UBE. In the remaining jurisdictions, the bar examination will continue relatively unchanged. Regardless of UBE status, NCBE will continue its mission of providing high-quality services to all bar examination authorities.

Issues/Questions for Reflection

- Today, "competency" and "fitness" are more typically two separate hurdles in legal licensure
- How do you predict the UBE will change over time?
- To what extent does the UBE improve the reliability and validity of state bar exams and better meet the needs of law schools with their national student bases and law school graduates with their multijurisdictional practices?

References

Albanese, M. A. (2013). *Cheating detection in professional education*. San Francisco, CA: Paper presented at the American Educational Research Association.

Albanese, M. A. (2014). Differences in subject area subscores on the MBE and other illusions. *The Bar Examiner, 83*(2), 26–31.

Albanese, M. A., Mejicano, G., Petty, E., Vuk, J., & McDonough, S. (2013). *Student cheating: Best practices for deterrence, detection, and disposition*. Madison, WI: Paper presented at the Conference on the Statistical Detection of Potential Test Fraud.

Albanese, M. A., & Ripkey, D. R. (2012). *Effects of rater monitoring on rater bias in grading essays on a high-stakes licensing examination*. Vancouver, British Columbia: Paper presented at the American Educational Research Association.

Albanese, M. A., & Wollack, J. (2013). *Cheating on tests: A threat to response validity*. Madison, WI: Paper presented at the Conference on the Statistical Detection of Potential Test Fraud.

American Bar Association. Council Statements. Retrieved from 21 May 2015, http://www.americanbar.org/content/dam/aba/migrated/legaled/accreditation/Council_Statements.authcheckdam.pdf

American Bar Association Center for Professional Responsibility. (2013). Model Rules of Professional Conduct. Retrieved from 20 May 2015, http://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/model_rules_of_professional_conduct_table_of_contents.html

American Bar Association (2010). Model Code of Judicial Conduct. Retrieved from 20 May 2015 http://www.americanbar.org/groups/professional_responsibility/publications/model_code_of_judicial_conduct.html

Bosse, D. F. (2011). The MPT: Assessment opportunities beyond the traditional essay. *The Bar Examiner, 80*(4), 17–21.

Case, S. M. (2004a). Decimal Dust. *The Bar Examiner, 73*(1), 33–34.

Case, S. M. (2004b). A quick guide to MPRE item development. *The Bar Examiner, 73*(3), 30–32.

Case, S. M. (2005a). Demystifying scaling to the MBE: How'd you do that? *The Bar Examiner, 74* (2), 45–46.

Case, S. M. (2005b). Feedback to candidates regarding MBE performance and an introduction to "Gain Scores". *The Bar Examiner, 74*(3), 28–29.

Case, S. M. (2006a). Frequently asked questions about scaling written test scores to the MBE. *The Bar Examiner, 75*(4), 42–44.

Case, S. M. (2006b). Men and women: Differences in performance on the MBE. *The Bar Examiner, 75*(2), 44–46.

Case, S. M. (2007a). How can we help? *The Bar Examiner, 76*(2), 41–42.

Case, S. M. (2007b). How to help repeaters improve their MBE scores. *The Bar Examiner, 76*(4), 42–44.

Case, S. M. (2008a). Best practices with weighting examination components. *The Bar Examiner, 77*(1), 43–46.

Case, S. M. (2008b). Relationships among bar examination component scores: do they measure anything different? *The Bar Examiner, 77*(3), 31–33.

Case, S. M. (2009). Coming together: The UBE. *The Bar Examiner, 78*(3), 28–33.

Case, S. M. (2010a). How to test knowledge of local law. *The Bar Examiner, 79*(3), 31–33.

Case, S. M. (2010b). Procedures for grading essays and performance tests. *The Bar Examiner, 79* (4), 36–38.

Case, S. M. (2010c). Top 10 list of best practices in testing for admission to the bar. *The Bar Examiner, 79*(2), 36–39.

Case, S. M. (2010d). The uniform bar examination: What's in it for me? *The Bar Examiner, 79*(1), 50–52.

Case, S. M. (2011). Common goals with increasingly similar outcomes: Jurisdiction approaches to bar exam grading, scoring, and standards. *The Bar Examiner, 80*(1), 53–55.

Early, K. R. (2011). The UBE: The policies behind the portability. *The Bar Examiner, 80*(3), 17–24.

Eckler, J. (1996). The multistate bar examination: Its origins and objectives. *The Bar Examiner, 65* (1), 14–18.

Gundersen, J. A. (2006). A new mix of questions on the multistate essay examination. *The Bar Examiner, 75*(3), 6–11.

Gundersen, J. A. (2007). Happy birthday, MPT. *The Bar Examiner, 76*(4), 18–25.

Hill, C. B. (2015). MBE test development: How questions are written, reviewed, and selected for test administrations. *The Bar Examiner, 84*(3), 23–28.

Kane, M. T. (2003). To round or to truncate? That is the question. *The Bar Examiner, 72*(4), 24–29.

Kane, M. T. (2009). Reflections on bar examining. *The Bar Examiner, 78*(4), 5–20.

Kane, M. T., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240.

Kane, M. T., & Mroch, A. A. (2005). Equating the MBE. *The Bar Examiner, 74*(3), 22–27.

Kane, M. T., Mroch, A. A., Ripkey, D. R., & Case, S. M. (2007). Pass rates and persistence on the New York bar examination including breakdowns for Racial/Ethnic Groups. *The Bar Examiner, 76*(4), 6–17.

Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2010). Linear equating for the NEAT design: Parameter substitution models and chained linear relationship models. *Measurement: Interdisciplinary Research and Perspectives, 7*, 125–146.

Karger, A. (1996). The continuing role of NCBE in the bar admissions process. *The Bar Examiner, 65*(2), 14–22.

Lee, S. Y., & Albanese, M. A. (2013). *An investigation of the detection of cheating on the multistate bar exam*. Madison, WI: Paper presented at the Conference on the Statistical Detection of Potential Test Fraud.

Lee, S. Y., Tracy, C., & Albanese, M. A. (2014). *A comparison of similarity indexes for detecting answer copying on the MBE*. Iowa City, IA: Paper presented at the 2014 Conference on Test Fraud.

Mroch, A. A., & Ripkey, D. R. (2007 April). *Structural Models Relating LSAT, Undergraduate GPAs, Law School GPAs, and Bar Examinations*. Chicago, IL: Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2010). An evaluation of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research and Perspectives, 7*, 174–193.

National Conference of Bar Examiners. About NCBE. *NCBE.* Retrieved from 19 May 2015, http://www.ncbex.org/about/

National Conference of Bar Examiners. Conference of Chief Justices Resolution 4–Endorsing Consideration of a Uniform Bar Examination. Retrieved from 4 Apr 2012, http://www.ncbex.org/assets/media_files/UBE/CCJ-Resolution-4-Uniform-Bar-Exam-2010-AM-Adopted.pdf

National Conference of Bar Examiners. Section of Legal Education and Admissions to the Bar Council Resolution Endorsing Consideration of a Uniform Bar Examination. Retrieved from 4 Apr 2012, http://www.ncbex.org/assets/media_files/UBE/ABA-Uniform-Bar-Exam-2010-Council-9-14-v2-3.pdf

National Conference of Bar Examiners. (2011a). 2012 MBE information booklet. Madison, WI.

National Conference of Bar Examiners. (2011b). 2012 MEE information booklet. Madison, WI.

National Conference of Bar Examiners. (2011c). 2012 MPRE information booklet. Madison, WI.

National Conference of Bar Examiners. (2011d). 2012 MPT information booklet. Madison, WI.

National Conference of Bar Examiners. (2015). Kansas adopts the Uniform Bar Examination (UBE). Retrieved from 20 May 2015, http://www.ncbex.org/news/kansas-adds-ube/

National Conference of Bar Examiners and American Bar Association. (2015). Comprehensive guide to bar admission requirements 2015. Madison, WI.

Pogers, J. (2010, December). Uniform bar examination is picking up steam. Retrieved 04/04/2012 2012, from http://www.abajournal.com/magazine/article/one_for_all_uniform_bar_exam_picking_up_steam/

Ripkey, D. R. (2008 March). Performance differences among ethnic groups: A national look within the legal education community. In S. M. Case (Ed.), *Annual meeting of the American Educational Research Association*. New York, NY.

Ripkey, D. R., & Case, S. M. (2009 April). *A cross-jurisdiction analysis of the relationships among scores on bar examination components*. San Diego, California: Paper presented at the Annual Meeting of the American Educational Research Association.

Ripkey, D. R., & Mroch, A. A. (April 2007). *The Reliability of Law School Grade Point Averages*. Chicago, IL: Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2010). An empirical comparison of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research and Perspectives, 7*, 147–173.

The New York State Board of Law Examiners. Uniform Bar Examination. Retrieved from 20 May 2015, http://www.nybarexam.org/UBE/UBE.html

Thiem, R. S. (2009). The Uniform Bar Examination: Change we can believe in. *The Bar Examiner, 78*(1), 12–14.

Tracy, C., & Albanese, M. A. (2014, September 30–October 2). *Disrupted opportunity analysis: A system for detecting unusual similarity between a suspected copier and a source.* Iowa City, IA: Paper presented at the Conference on Test Fraud.

VandeWalle, G. W. (2009). Life without a local bar exam. *The Bar Examiner, 78*(1), 7–9.

Chapter 21
Summary and Conclusions

Paul F. Wimmers and Marcia Mentkowski

Abstract There is a growing skill gap between students' level of preparedness after finishing school and demands from the workforce. During the past decades, the transformation from industrial to postindustrial economies has changed the context of young graduate's transition to the labor market. In most western countries, these transitions have become not only prolonged but also more fragmented, diversified, and less linear. As a result, employers have difficulty in finding appropriate candidates while many graduates cannot find jobs. Youth unemployment figures are often twice as high as those for adults.

There is a growing skill gap between students' level of preparedness after finishing school and demands from the workforce. During the past decades, the transformation from industrial to postindustrial economies has changed the context of young graduate's transition to the labor market. In most western countries, these transitions have become not only prolonged but also more fragmented, diversified, and less linear. As a result, employers have difficulty in finding appropriate candidates while many graduates cannot find jobs. Youth unemployment figures are often twice as high as those for adults. There is a noticeable mismatch between outdated qualifications and the rapidly changing demands of the labor market (Walther and Plug 2006; Walter and Markus 2003). Questions about the employment value of college or university degrees have intensified. Program directors of medical residencies, for example, have increasingly expressed concerns that graduates are not well prepared for residencies (Lyss-Lerman et al. 2009). Where is this misalignment coming from? Most educators will agree that educating professionals requires formal education. Proper education and training is a key resource on entry into working life. It is essential that, faculty members, administrators and educators

P.F. Wimmers (✉)
David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
e-mail: pwimmers@ucla.edu

M. Mentkowski
Alverno College, Milwaukee, USA

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1_21

are well-prepared to face these new challenges and are provided with authentic assessment strategies and technologies that offer support. This book is a start.

This final chapter of the book will give a summary of all chapters and will discuss the common themes and findings.

21.1 Knowledge, Performance and Competence

Being knowledgeable has always been valued in our modern society. In the late nineties and early twenty-first century, there was a global focus on the need for a knowledge-based society (Garnham 2002), (Russell, McKee, Westhead, Chap. 9). There are many who appear to believe that this knowledge is all that needs to be assessed and it is unquestionable measurements of knowledge that dominates current institutional and specialty board examination systems (Miller 1990). But knowledge is only valuable when it can be applied in practice (Whitehead 1929). Testing of what students know, will not determine if students know how to apply this knowledge. When and how do we know if students have the ability to do something successfully? White (1959) suggested that competence or the ability to do something successfully, be considered a complex construct and conceptualized it as effective interaction with the environment. This means that competence as a construct or quality manifests itself only in observed behaviors or practice and not in a written test. Miller (1990) said competence was the measurement of an examinee's ability to use his or her knowledge, this to include such things as the acquisition of information, the analysis and interpretation of data, and the management of problems.

A short discussion of a framework for assessment proposed by Miller (1990) clarifies the relationship between knowledge performance and competence (see Fig. 21.1).

In his framework for assessment he made a distinction between knows (knowledge), knows how (competence), shows how (performance) and does

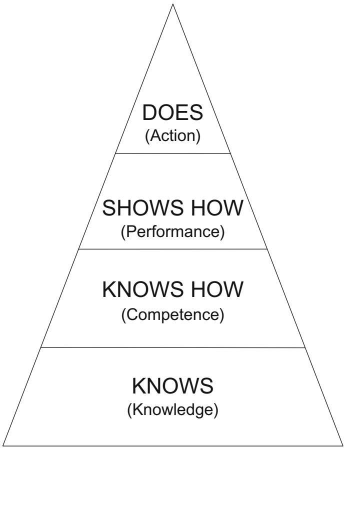


Fig. 21.1 Framework for assessment by Miller (1990)

(action). Knowledge is at the base of this triangle shaped framework and action is at the top. A student, resident, or physician needs the knowledge that is required to carry out professional functions effectively, a prerequisite for being competent. The next two layers, competence and performance, which follow upon knowledge, are often used interchangeably; however, competence means that a physician can apply his/her knowledge in concrete situations, while performance is the ability to use this knowledge to perform concrete actions. The final top layer represents what a professional actually does during day-to-day practice.

A distinction between competence and performance is often made in the literature. Senior (1976), for example, defined competence as what a physician is capable of doing and performance what a physician actually does. The former would, in this case, be related to the first three layers of the pyramid of Miller (1990) and the latter to the top layer (See also, Rethans et al. 1990; Van der Vleuten and Swanson 1990). The implication is that performance must be measured or observed in order to assess competence, and many different tests are probably needed. So we are not concerned with performance per se, "rather we are concerned about conclusions we can draw and the predictions we can make on the basis of that performance" (See also Chap. 5, Heywood).

Assessment has focused mostly on "knows" and "knows how," the base of the pyramid: recall of factual knowledge and the application of this knowledge in problem solving. However, such examinations may fail to document what students will do when faced with real work situations. To determine someone's competence, observing behaviors in action is needed, and this is represented by the top layer of the pyramid in Miller's model (Miller 1990; Wass et al. 2001) Despite the fact that Miller's pyramid is primarily intended to serve as a framework to define and categorize different assessment tools, his model gives a good idea about which characteristics are influencing the development of competence. The different layers in Miller's model represent a developmental sequence of stages in the process from novice to expert. The layers represent how students build their knowledge during their initial years of training and how competence and performance are shaped in action during the latter years in apprenticeship and practice.

21.2 Assessment of Outcomes and Competencies

Traditional knowledge-based assessments appeared to have low predictive success for professionals upon graduation (Darling-Hammond and Snyder 2000), and did not seem to reinforce the types of knowledge needed to succeed in the professions (Gibbons et al. 1994), (Chap. 4, O'Neil). To serve the needs of the professional workplace we need to focus on students' ability to use his or her knowledge, and competence was the measurement of an examinee's ability to use his or her knowledge. Competency-based assessment is the assessment of a person's competence against prescribed standards of performance. Thus, if a profession has established a set of competency standards, then these prescribe the standards of performance required of all new entrants to that occupation. Competency-based assessment is the process determining whether a candidate meets the prescribed standards of performance, i.e. whether they demonstrate competence (Hagera et al. 1994). Testing against a standard is not new and is historically called a criterion-referenced test. However, assessment of outcomes or competencies in professional education gained popularity in the early 2000s. The move toward competency-based educational programs started around 2000 in medicine when the Accreditation Council for Graduate Medical Education (ACGME) decided to specify six general competencies of graduate medical education (Chap. 19, Park, Hodges, and Tekian). By 2000 the engineering curriculum had come to be dominated by outcome approaches to assessment (Chap. 5, Heywood). There is an overwhelming necessity to improve the alignment between learning outcomes, assessment, and demands of the workforce. Professional schools often lag in adapting to new challenges while rising expectations characterize learning in the workplace. Well-defined professional outcomes are key in bridging the gap between education and the workforce. However, learning outcomes are often complex and multidimensional and performance assessment takes place at the interaction of the integrating of subject matter knowledge and learned abilities. Faculty members may require students not only to practice but demonstrate, adapt, and transfer learned capacities, defined as integration of learned abilities with patterns of performance (Chap. 2, Mentkowski, Diez, Lieberman, Pointer Mace, Rauschenberger, and Abromeit).

Harris in Chap. 3 defines curriculum broadly and sees assessment and learning outcomes as part of the curriculum. Assessment of performance is an essential component of any curriculum, across the professions. Assessment of performance in the actual settings of practice, or in authentic simulations, is essential. “Curriculum” does not refer solely to the content or subject matter of an education program. Consistent with current conceptions of curriculum practice and scholarship, “curriculum” refers more broadly to every facet of the planning and implementation of education programs, including: general and targeted needs assessment; formulation of learning goals and objectives; selection of approaches and methods of instruction, including the teaching and learning environment; assessment of learners; and evaluation of the education program (Shubert et al. 2002; Reynolds et al. 1996)

There are some issues with the assessment of competencies that should not be overlooked. Competencies are (1) context dependent, (2) multi-dimensional and interconnected, and (3) sensitive to time. Heywood in Chap. 5 mentions that the belief that students can be prepared for work immediately on graduation by the acquisition of specifically stated competencies that can be taught has now been challenged on several occasions. A phenomenological study of engineers of work is reported by Sandberg (2000) that offers an alternative view of competency furthers this view. Competency is found to be context dependent and a function of the meaning that work has for the individual involved. Engineers were found to have different perceptions of work, and competencies related to the same task were found to be hierarchically ordered among them, each level being more comprehensive than the previous level. Attributes are developed as a function of work. It follows that they are not fixed, therefore firms will have to undertake training (or professional development) beginning with an understanding of the conception that the engineers has of her/his work. Professional competence should be regarded as reflection in action or understanding of work or practice rather than as a body of scientific knowledge.

The Chap. 6 of Wimmers, Guerrero, and Baillie (6) addresses the question "how residents perceive they acquire proficiency in the core competencies." Various authors have pointed to the difficulty in knowing how well residents have acquired a competency and how these can be effectively taught (Caverzagie et al. 2008; Cogbill et al. 2005; Lurie et al. 2009). Core competencies and related learning objectives are considered educational outcomes and medical residents are required to demonstrate sufficient proficiency in all of these competencies independent of their residency. This means that professional training is primarily driven by output measures (objectives, competencies) rather than input measures (instruction, learning activities). There is assumed that assessment based upon the core competencies provides evidence of the program's effectiveness in preparing residents for practice. Specific educational activities foster multiple competencies. Competencies are not perceived to be learned through any single learning activity. The fact that competencies are multi-dimensional and interconnected makes it highly unlikely that a single approach to teaching or assessment will be sufficient for their acquisition and multiple methods for teaching and learning are necessary for the acquisition of the competencies. Clinical competence takes place at the intersection of a lot of different learned abilities and skills. This ability of implementing and applying multiple core competencies is what medicine is about. Instruction and assessment are very closely related although they seem different (See Chap. 6, Wimmers et al.).

Another important consequence of the transition toward competency-based programs is timing. This issue is discussed in Chap. 19 of Park et al. The intent of competency-based programs is that only students who are judged "competent" would be able to move forward in a professional school curriculum. Two fundamental issues arise with the fixed, time-based model. First, assessment of competence is a challenge. Courses are structured into pre-determined intervals, making ongoing, meaningful feedback structurally difficult in this curricular setting. While summative assessments are common in this setting, they do not necessarily provide information about whether learners have acquired the competencies to perform at work. Implementing workplace-based assessments that allow continuous measurement of skills in the clinical workplace can be difficult in a time-based environment, because of a lack of alignment between instruction in time-based modules and assessment of overall educational outcomes. Within this context, Holmboe et al. (2010) outlined six components of an effective assessment system in CBME:

1. Assessments need to be continuous and frequent
2. Assessments must be criterion-based, using a developmental perspective
3. Competency-based medical education requires robust work-based assessment
4. Training programs must use assessment tools that meet minimum standards of quality

5. We must be willing to incorporate more "qualitative" approaches to assessment
6. Assessment needs to draw upon the wisdom of a group and to involve active engagement by the trainee

21.3 Authentic Assessment

Authenticity has not always been a given in professional education assessment. During much of the 80s and 90s, a fierce argument raged over the relative value of authenticity in assessment. O'Neil, in Chap. 4, reminds us that the most authentic assessments occur not in school or training programs, but in the workplace. Because of the considerable overlap between the classroom and actual practice in the professions, professional educators would be wise to broaden their view of assessment beyond that of a practice that stops upon graduation. He suggests three tools for assessing performance in the medical workplace: outcomes measures (e.g. patient morbidity and mortality), large-scale data collection and processing (e.g. chart review), and portfolios. However, O'Neil concludes, there is no reason that these three tools, and others, couldn't be implemented earlier in a professionals education.

Harris in Chap. 20 also emphasizes early workplace learning and assessment. Early learning in the workplace provides essential experiences that are consistent with developing the competencies needed for professional practice, given the nature of professional practice (Schon 1987). Clearly, each profession requires specialized and often sub-specialist knowledge. But in addition, professional practice situations are characterized by conditions of complexity, uniqueness, uncertainty, ambiguity, and conflicting value orientations (Harris 2011). In turn, effective professional practice requires 'practical knowledge' for applying a repertoire of specialized knowledge in various specific situations and related reflective competencies for self-assessment, responding to others' assessments, independent learning and self-correction, typically learning in practice, through experience with the actual problems of practice. In education for the professions, learning in the practice settings of the workplace is the signature pedagogy, where novices are socialized into a community of practice; develop professional skills through observation, role modeling, practice, reflection and feedback; and develop the motivation and context for application of knowledge learned in classroom settings (Harris 2011).

21.4 Interdisciplinary and Interprofessional Assessment

The world is becoming increasingly complex and demanding, with further specialization needed in professional development. Specialists and sub-specialists, who represent a narrower field of study within a discipline, have created their own unique definitions, acronyms, and terms. Yet the need for engagement in learning and human interactions increases in a complex and global world. Communication and collaboration among individuals of different professions is becoming even more challenging. This is not only obvious on the work floor of large engineering and construction projects but also in managing complex rescue operations, launching a space shuttle or producing big-budget movies. Complex organizational environments require multiple areas of expertise and have to develop the ability to work effectively with diverse stakeholders. Learning in the professions is best understood as a process embedded in social relationships and social practices, and other professionals participate in these relationships and practices over time and across settings (Curry and Wergin 1993; Peck et al. 2010; Wenger 1998). In medical care, for example, multidisciplinary teams are increasingly used for diagnosis and discussion of complicated treatment options and their outcomes. The opinions of an individual physician are making a place for higher-order group decisions. An oncologist with a pancreatic cancer patient has to work together with surgeons, radiologists, palliative care physicians, nurses, dietitians, and hospital administrators. A psychiatrist involved in the assessment of an abuse case may work with professionals from other disciplines such as protective services or civil or criminal justice specialists.

As reliance on teams in organizations increases, team training and evaluation of team performance becomes more important. In many situations, a well-functioning team can accomplish more than the sum of its individual parts (Doyle et al. 2013). Evaluation and assessment of team performance should focus on both, the performance of an individual in a team and the performance of the team as a whole. Team training starts in professional schools and many disciplines make use of simulated-based training for teams (Webb 1980, 1982).

Interaction among professionals in a multidisciplinary environment can take different forms and often different descriptions are used in the literature: (1) multi-disciplinary; (2) cross-disciplinary/cross-professional; (3) interdisciplinary; or (4) trans-disciplinary. Problem-solving in an interprofessional team context takes on meaning depending on the professions represented. Different professions may apply similar problem-solving techniques, but they solve very different problems and hence their performances are diverse. Another goal for graduates is that they become able to *translate* their profession for other professionals without engaging in useless semantics and value conflicts. Rather, they can practice contributing to identifying, clarifying, and resolving some of the great problems of our time. These problems include arguing persuasively in community settings, and generating solutions in group meetings by both brainstorming and useful critique (Isaksen 1998). Serving on panels of professionals to represent their own profession is a common occurrence. Yet, capturing and clarifying the ideas of another colleague to build on his or her ideas uses civil discourse, rather than engaging in competition promoted by either or questions so familiar in the media. Ultimately, we expect graduates to participate in developing sustainable policies in all professions.

McKinley in Chap. 14 reviews research conducted with healthcare professionals to determine the extent to which assessments of team performance had been developed and evaluated between 2006 and 2012. The nature of the ‘hypercomplex environment’ in which health care occurs is characterized by several decision makers, whose roles are embedded in an 'extreme hierarchical differentiation', they note that the assurance of patient safety requires interaction and communication in 'compressed time' with a 'high degree of accountability' (Baker et al. 2006). In identifying the characteristics of high-reliability organizations, Baker et al. (2006) argue that healthcare providers are often organized in teams, and that their interactions are part of the vital operations in various settings. The hypercomplexity of the context in which health care occurs is characterized by specialization, where team members have specific roles, responsibilities and knowledge (Orchard et al. 2012). Because errors, although rare, result in serious consequences, teamwork is essential. Knowledge of their own roles and responsibilities, monitoring of team member performance, and a positive attitude towards teamwork have been shown to relate to team effectiveness. Team competencies typically considered for training programs have been identified as leadership. Mutual performance monitoring, mutual support, adaptability, team orientation, mutual trust, shared mental models, and communication (Baker et al. 2006).

Teams are as diverse as the communities they serve. A focus on improving coordination and communication between departments is of life importance for the future of health care (Frenk et al. 2010; Lamb et al. 2011; Ruhstaller et al. 2006; Tattersall 2006). Employers view the ability to collaborate with others as a core twenty-first century competency that is more important than even English language ability or subject matter knowledge for both landing and keeping a job. Many of today's professional problems require interprofessional solutions. Successful students may be better prepared to succeed in undergraduate professions when they demonstrate integration of their knowledge systems and competencies learned and assessed in math and science courses. However, whether students are able to adapt and transfer what they have learned on demand in assessments that require them to demonstrate new uses with unfamiliar problems that require analysis and problem solving is a question for faculty across higher education (Chap. 8: Mentkowski).

The purpose of Webb's Chap. 13 is to enumerate and describe the challenges that emerge in performance assessments that include groupwork. Her chapter focuses on performance assessments in which individuals work together in small groups of size two or more to achieve a common goal, whether it is solving a problem, completing a task or activity, or producing a product. While many performance assessments require only individual performance, they can be extended to involve individuals working in groups on a common task or set of tasks to reflect the value that educators, policy makers, the business community, and the general public place on the ability to communicate and collaborate effectively with others (Baron 1992). In addition to the usual challenges of measuring examinees' performance on complex tasks carried out individually, groups working on common tasks present unique or more complicated measurement challenges. Common themes running through the many taxonomies and frameworks of teamwork and teamwork skills include adaptability (recognizing problems and responding appropriately), communication (exchanging clear and accurate information), coordination (organizing team activities to complete a task on time), decision-making (using available information to make decisions), interpersonal (interacting cooperatively with other team members), and leadership (providing structure and direction for the team, Chung et al. 1999; SCANS 1999). By definition, these collaboration, communication, and teamwork skills involve interaction with others. Incorporating groupwork into assessments provides a direct way of measuring these skills as well as the productivity and performance of the group as a whole.

Webb's Chap. 13 also describes the many sources of variation that can impact the reliability and validity of performance assessment in groupwork situations. The nature of group processes that arise in a particular groupwork session may greatly impact scores of groups and/or their members. Several ways in which group functioning may differ, with consequences for assessment scores, include: task-related interaction with others, lack of involvement, uncoordinated group communication, social-emotional processes, or the division of labor. The group work setting also introduces new sources of score variability that do not pertain to individual assessments, such as the composition of the group and the roles played by group members, the type of the task, type of occasion, type of rater, and type of rating scale used.

Simmons, Wagner, and Reeves in Chap. 12 discuss key issues in assessment of interprofessional education (IPE). IPE can have a beneficial impact on learners' ability to work together in an effective manner collaborative (Hammick et al. 2007; Reeves et al. 2013; Zwarenstein et al. 1999). IPE aims to provide learners with interactive experiences in order to better prepare them working collaboratively to effectively meet the demands of the task. While the evaluation of IPE programs continues to grow, in contrast the assessment of learning in IPE has received far less consideration, with only a limited amount of literature published. Key questions to consider in relation to assessing IPE include: should one use a summative approach to assessing learning in interprofessional groups or teams or is a formative assessment approach more effective or should both be utilized? What should be the focus of the assessment: the individual, the team or the completion of the task? What should one measure: individual-based, patient/client centered-based, organizational-based outcomes or others? The competencies for IPE are domain independent outcomes and are related to each level of Miller's typology and independent of content (Miller 1990). The authors of this chapter suggest using a multi-factor approach to assessment by examining team structure (made up of individuals), the functions of the team (understanding their roles, responsibilities and relationships) and outcomes (task completion).

In Heywood's Chap. 5 it has been argued that teamwork can contribute to the development of innovation skills and creativity. The research reported on the former suggested that heterogeneous teams were not more innovative than homogenous teams (Fila et al. 2011). It has been found that high levels of interdisciplinary and integration may contribute to positive learning experiences (Froyd and Ohland 2005). However, it is suggested that many students are challenged by collaboration skills. Such skills have a large affective component and are context dependent as a function of an individual's personality. Communication skills are particularly challenged when groups have different perceptions of the problem. Transdisciplinary projects are able to integrate the tools, techniques and methods from a variety of disciplines. Impediments to collaboration include disciplinary prejudice, unwillingness to listen and ask questions, and lack of shared ideas (Lingard and Barkataki 2011). A key problem that is not fully understood is the level of knowledge required by each partner in the other disciplines involved. “Constructive Controversy” has been recommended as means of creating mutual understanding about a problem area (Johnson and Johnson 2007; Matusovich and Smith 2009). An experimental course based on constructive controversy led to the reminder that the pedagogic reasoning for the use of non-traditional methods of instruction needs to be explained (Daniels and Cajander 2010).

McKinley’s Chap. 14 lists a summary of challenges that the measurement of teamwork amongst healthcare professionals faces. First, efforts continue to be specialty-specific (e.g., surgical, emergency medicine, community medicine), although there are studies that have looked to see if the measures can be used across setting (e.g., O’Leary et al. 2012). While several of the measures developed are based in theory, different constructs may be measured. Although there was minimal inconsistency in terminology, papers that do not clearly define the constructs measured can make this effort challenging, particularly if measures are to be used across health professionals and settings.

Finding reliable measurement tools that assess group interaction can be a challenge particularly when targeted training in teamwork skills has been conducted. Research has shown that team members are generally not reliable at assessing their level of skill (Baker and Salas 1992; Eva et al. 2004), but practitioners are generally able to self-monitor (Eva and Regehr 2011). Although observational measures have been said to be preferable, securing the necessary number of raters to produce reliable measures has been challenging (Morgan et al. 2007) although recent work has shown promise (Russ et al. 2012). Efforts are underway to show that shorter version of long measures can be used in a fashion that may facilitate recruitment and training of raters, generating more ratings available for the evaluation of teamwork skills (Lurie et al. 2011).

McKee’s Chap. 12 and Lee’s Chap. 17 are both using an action research approach. Action research, a systematic inquiry process that engages the participants in a series of “learning by doing” activities to achieve the goal of organizational changes and quality improvements, was widely popular in the business world during the 1980s and has since been vigorously adopted in the educational environment (French and Bell 1995; Mills 2000). McKee’s Chapter uses an action research approach to help primary care teams based in general practice to develop a productive culture of work-based learning and reflective practice, thus enriching the learning environment within practices. Key characteristics of the workplace of primary care were identified and these suggested the need to re-think assessment-for-improvement that is team and organizationally-based. The study identifies: (1) The impact of policy on practice, learning, assessment, and accountability when practicalities matter. (2) Complexities of administering comparable assessments of work-based learning when stakeholders and primary care professionals interpret project purposes and outcomes differently. (3) Challenges when developing practitioner-conducted assessments of learning arising from everyday practice where heavy workloads couple with high external demands.

Lee's Chapter uses an action research approach to identify important themes in training in leadership. Leaders in all professions are expected to possess key leadership skills, such as teamwork, communication, consensus building, conflict resolution, and forward-looking vision, to cope with rapid and profound changes in an environment rife with financial, ethical, and profession-specific complexity (Kuo et al. 2010; Ackerly et al. 2011). The purposes of this study are to conduct both quantitative and qualitative analyses of the action research reports submitted by the Leadership and Management in Geriatrics program participants to assess their performance and experience in implementing action research, and to examine the value of action research as an extended educational mechanism in leadership training. Four categories of themes were identified: Action Planning, Implementation Process, Outcome and Impact, and Follow-up Activity.

21.5 Performance Assessment in Legal Education

In Chap. 7, Abner and Kierstead provide a case study of performance assessment at Osgoode Hall Law School. The curriculum emphasizes self-reflection and self-assessment, direct professional practice, and experiential learning. Although the curriculum does include summative assessment components, much of the focus is on formative assessment. The role of uncertainty in legal professionalism is drawn from Mentkowski et al's outcomes framework in legal education to underscore the importance of professional judgment under conditions of uncertainty. The importance of professional judgment under uncertain conditions is reflected in the Multistate Performance Test as well, where examinees must "separate relevant from irrelevant facts," and perform adequately despite the fact that facts presented within the test "are sometimes ambiguous, incomplete, or even conflicting [and] a client's or a supervising attorney's version of events may be incomplete or unreliable." Another aspect is ethics. Their program has ethical lawyering requirements, including dedicated class time devoted to ethical decision making, the incorporation of ethical considerations across all courses in the law program (per a Faculty Council directive), and reflective exercises inviting students to make connections between their formal ethics training and their public interest work experiences.

Chapter 20, Ripkey and Kane describe the development of a large-scale summative assessment: the Uniform Bar Examination. This high-stakes licensure examination is offered twice annually, and passage is required for legal practice in the United States and a handful of other jurisdictions. Ripkey and Kane outline the challenges and rewards of a move towards standardization of the Bar Examination as experienced by various stakeholder groups. The National Conference of Bar Examiners produces a stand-alone Multistate Professional Responsibility Examination required by most US jurisdictions for admission to the practice of law.

Within the context of the Uniform Bar Examination too, examinees are required to "identify and resolve ethical dilemmas" as part of the Multistate Performance Test.

Thus, although the two chapters focusing on legal education are different from each other in many ways, they do share some points in common: they both highlight trends in legal education towards performance and professional ethics (with neither losing sight of the importance of substantive knowledge of the law as well for competent legal practice.) They also both include a focus on complexity and uncertainty in professional judgment. Both advocate the use of frameworks—in developing curriculum (i.e., both formative and summative assessments) and in licensing examinations. The reliance on these frameworks focuses the development of test materials, and tying performance assessment to job analyses lends validity evidence to curricula, tests and testing programs. Finally, in keeping with a major theme of this book, both emphasize the importance of collaboration across stakeholders and groups, despite the fact that this collaboration can be challenging at times.[1]

21.6 Competencies that Cross the Disciplines

Rencic, Durning, Holmboe, and Gruppen's Chap. 11 reviews the concept of "reasoning," in particular, clinical reasoning. They conclude that there is a strong desire among educators to assist learners in developing good clinical reasoning, and the need for meaningful and effective remediation for those struggling with clinical reasoning. Reflecting on the literature, they believe that the strongest recommendation that can be made is for educators to focus on helping learners build their discipline-specific knowledge and its organization (Elstein 1972). Given that much of the clinical reasoning process can be subconscious and is idiosyncratic (i.e., two health care professionals may come to the same conclusion using different processes based on their knowledge and experiences), educators must recognize that no "gold-standard" for the clinical reasoning process exists. In this relativistic world, knowledge assessment can provide a foundation. When a learner misses a diagnosis, the focus can first turn to the gaps in her knowledge (i.e., what knowledge was faulty or lacking that led her to the wrong diagnosis?).

Fountain in Chap. 10 reviews the concept "critical thinking." Critical thinking is a key competency for health professionals to function in our complex health care environment. There is far less clarity in professions education literature about what it really means or how to best measure it in providers and students. Fountain concluded that an explicit definition was not provided in almost half of the studies. Keyword analysis identified 6 common constructs in nursing studies of critical thinking in nursing: individual interest, knowledge, relational reasoning, prioritization, inference, and evaluation, along with 3 contextual factors: patient

[1] Special thanks to Joanne Kane, co-author of Chap 20, for highlighting the commonalties between the two legal chapters.

assessment, caring, and environmental resource assessment. Fountain concluded that critical thinking research needs to use explicit definitions that are part and parcel of the measures that operationalize the construct, and alternative measures need to be developed that line up the attributes of the definition with the attributes measured in the instrument. Once we can accurately describe and measure critical thinking, we can better ensure that rising members of the professions can apply this vital competency to patient care.

Lee, Wimmers, and Fung's Chap. 18 embraces the concept "humanism," and discusses the development of an instrument to measure humanism. Physicians are expected to demonstrate not only clinical competencies but also caring attitudes and behaviors. Even though this chapter focuses on humanism in patient care, humanism is not only considered the core belief and value of medical professionals but many other professions were the human relationship and human interaction is central. We saw for example in Chap. 15 (Abner and Kierstead) that law school's program has ethical lawyering requirements, including ethical decision making. A review of the literature in medical humanism led to the definition of a framework for humanism with five personal attributes: integrity, compassion, altruism, respect, and empathy. These attributes are observable in patient-centered care behaviors and attitudes. The humanism instrument was validated by scoring clinical performance examination video recordings of student interactions with patients.

The fact that formal training and assessment of concepts like humanism in authentic workplace settings is so important is echoed in Chap. 3. Harris addresses the importance of the relationships between curricula and the cultural, social, political, and economic structures of the professional school and workplace setting; the hidden curriculum of role modeling and professional socialization; and the curriculum that students actually experience. Studies in the role of the hidden curriculum in development of health professionals in both academic institutional settings and the practice settings of the workplace almost uniformly demonstrate that the professional values recommended in the formal curriculum, are not in fact, consistently demonstrated in the practice settings of the workplace. For example, Stern (1998) reports a study comparing the "recommended curriculum" of medical values, identified through content analysis of curriculum documents, with the values actually taught, in hospital-based internal medicine teams at an academic medical center, identified through naturalistic, but systematic, observation. Among his findings was that the while the formal curriculum emphasized the importance of inter-professional respect, the naturalistic observation of actual practice in the setting of the workplace, demonstrated pervasive professional disrespect (Stern 1998).

21.7 Conclusions

Learning in the professions is about preparing students to work in the profession. To align what is learned with what is needed in the workplace; assessment should be about measuring performance. So assessment in the professions is not about asking what you have learned, but about showing how you apply what you have learned in a realistic or authentic context. We do not mean to say that knowing how, and specific content knowledge about your profession, is not important. It surely is, but to be confident that your curriculum has prepared your student to do the job, there is only one question that needs to be answered: How competent is your graduate to successfully function in his profession? Is it realistic to demand from graduates to be prepared for the workplace? We think it is. We, as educators, cannot demand a high commitment, financially and emotionally, from our students while doing a mediocre job in preparing them. A big step forward in aligning the curriculum with the demands of the workplace is the definition of outcomes and competencies. Competencies are the end goal of the curriculum and should closely match the competencies needed in the workplace. But competencies cannot only be defined on the individual level. We saw the importance of teamwork and group work. The reliance on teams in organizations is steadily increasing; team training and the evaluation of team performance is becoming more important than ever

Table 21.1 Key concepts for assessing competence in professional performance

Self assessment: Professional schools are now expected to graduate students who can evaluate their own work, continuously improve their performance, and engage in lifelong learning
Authenticity of assessment: In contrast to conventional multiple-choice tests, performance assessments require examinees to respond to complex tasks that represent authentic or "real world" problem-solving or performance situations and, ideally, do a better job in assessing examinees' higher-order thinking skills, deep understanding of content, complex problem solving, and communication
Early professional experiences and early authentic assessment: Should professional schools prepare for work? We think they do. One way in accomplishing this challenge is to integrate knowledge within the application of practice
Communication and collaboration among individuals of different professions: The sub-specialization of professions will require better communication and collaboration to maintain a holistic perspective on care and service
Assessment of teamwork: To assess individuals' capabilities in working with others to accomplish a common task. It is a measurement of communication, collaboration and teamwork skills
Professional judgment under uncertain conditions: Students need to be able to separate relevant from irrelevant facts quickly and maintain the ability to perform adequately despite the fact that facts presented are often ambiguous, incomplete, or even conflicting with each other
Professional development: The curriculum is delivered by our teachers, faculty and staff. Proper guidance and professional training in teaching and assessment is extremely important for any professional school

before. Teams communicate and collaborate across disciplines and professions and the definition of team competencies and the proper assessment of team competencies should include but also go beyond the performance of an individual within that team. See Table 21.1 for an overview of key concepts.

We would like to end this last chapter by reiterating what Dr. Shulman said in his foreword. We have to think of the development of professional learning as "learning to profess." And this kind of learning goes beyond the cognitive, it comprised of three distinctive yet interacting kinds of learning: (1) the development of habits of mind, (2) habits of practice, and (3) habits of the heart.

References

Ackerly, D. C., Sangvai, D. G., Udayakumar, K., Shah, B. R., Kalman, N. S., & Cho, A. H. (2011). Training the next generation of physician-executives: An innovative residency pathway in management and leadership. *Academic Medicine, 86*, 575–579.

Baker, D. P., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research, 41*(4 Pt 2), 1576–1598. doi:10.1111/j.1475-6773.2006.00566.x

Baker, D. P., & Salas, E. (1992). Principles for measuring teamwork skills. *Human Factors, 34*, 469–475.

Baron, J. B. (1992). *SEA usage of alternative assessment: The connecticut experience*. Paper presented at the Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd ed.). Washington, DC, September 4–6, 1991.

Caverzagie, K., Shea, J., & Kogan, J. (2008). Resident identification of learning objectives after performing self-assessment based upon the ACGME Core Competencies. *Journal of General Internal Medicine, 23*(7), 1024–1027. doi:10.1007/s11606-008-0571-7

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15*, 463–493.

Cogbill, K. K., O'Sullivan, P. S., & Clardy, J. (2005). Residents' perception of effectiveness of twelve evaluation methods for measuring competency. *Academic Psychiatry, 29*, 76–81.

Curry, L., & Wergin, J. F. (Eds.). (1993). *Educating professionals: Responding to new expectations for competence and accountability*. San Francisco: Jossey-Brass.

Daniels, M., & Cajander, A. (2010). *Experiences from using constructive controversy in an open ended group project*. Paper presented at the Proceedings Frontiers in Education Conference, ASEE/IEEE. S3D-1 to 5.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and teacher education, 16*(5), 523–545.

Doyle, C., Wilkerson, L., & Wimmers, P. F. (2013). Clinical clerkship timing revisited: Support for non-uniform sequencing. *Medical Teacher, 35*(7), 586–590. doi:10.3109/0142159X.2013.778393

Elstein, A. S. (1972). *Methods and theory in the study of medical inquiry*. Cambridge, MA: Harvard University Press.

Eva, K. W., Cunnington, J. P., Reiter, H. I., Keane, D. R., & Norman, G. R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Advances in Health Sciences Education, 9*(3), 211–224.

Eva, K., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3), 311–329. doi:10.1007/s10459-010-9263-2

Fila, N. D., Wertz, R. E., & Purzer, S. (2011). *Does diversity in novice teams lead to greater innovation?* Paper presented at the Proceedings Frontiers in Education Conference, ASEE/IEEE. S3H-1 to 5.

French, W., & Bell, C. (1995). *Organization development: Behavioral science interventions for organization development*. Englewood Cliffs, NJ: Prentice-Hall.

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., & Zurayk, H. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet, 376*(9756), 1923–1958.

Froyd, J. E., & Ohland, M. W. (2005). Integrated engineering curricula. *Journal of Engineering Education, 94*(1), 147–164.

Garnham, N. (2002). Information society's theory or ideology: A critical perspective on technology, education and employment in the information age. In W. H. Dutton & B. D. Loader (Eds.), *Digital Academe. The new media and institutions of higher education and learning*. London: Routledge.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage.

Hagera, P., Gonczia, A., & Athanasoua, J. (1994). General Issues about assessment of competence. *Assessment & Evaluation in Higher Education, 19*(1), 3–16.

Hammick, M., Freeth, D., Koppel, I., Reeves, S., & Barr, H. (2007). A best evidence systematic review of interprofessional education. *Medical Teacher, 29*, 735–751.

Harris, I. (2011). Conceptual perspectives and the formal curriculum. In J. P. Hafler (Ed.), *Extraordinary learning in the workplace*. Dordrecht: Springer.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*, 676–682.

Isaksen, S. G. (1998). *A review of brainstorming research: Six critical issues for enquiry (Monograph #302)*. Buffalo, NY: Creative Problem Solving Group-Buffalo.

Johnson, D., & Johnson, R. (2007). *Creative Constructive Controversy. Intellectual Challenge in the Classrooms* (4th ed.). Edina Min: Interaction pbl.

Kuo, A. K., Thyne, S. M., Chen, H. C., West, D. C., & Kamei, R. K. (2010). An innovative residency program designed to develop leaders to improve the health of children. *Academic Medicine, 85*, 1603–1608.

Lamb, B. W., Sevdalis, N., Arora, S., Pinto, A., Vincent, C., & Green, J. S. (2011). Teamwork and team decision-making at multidisciplinary cancer conferences: barriers, facilitators, and opportunities for improvement. *World Journal of Surgery, 35*(9), 1970–1976.

Lingard, R., & Barkataki, A. (2011). *Teaching teamwork in engineering and computer science*. Paper presented at the Proceedings Frontiers in Education Conference, ASEE/IEEE. F1|C-1 to 5.

Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2009). Measurement of the general competencies of the accreditation council for graduate medical education: A systematic review. *Academic Medicine, 84*(3), 301-309. doi:10.1097/ACM.1090b1013e3181971f3181908

Lurie, S. J., Schultz, S. H., & Lamanna, G. (2011). Assessing teamwork: a reliable five-question survey. *Family Medicine, 43*(10), 731–734.

Lyss-Lerman, P. M., Teherani, A., Aagaard, E., Loeser, H., Cooke, M., & Harper, G. M. (2009). What training is needed in the fourth year of medical school? Views of residency program directors. *Academic Medicine, 84*(7), 823–829.

Matusovich, H., & Smith, K. (2009). *Constructive academic controversy-what is it? Why use it? How to structure it*. Paper presented at the Proceedings Frontiers in Education Conference, M3A—1 to 3.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9 Suppl), S63–S67.

Mills, G. (2000). *Action research: A guide for the teacher researcher*.

Morgan, P. J., Pittini, R., Regehr, G., Marrs, C., & Haley, M. F. (2007). Evaluating teamwork in a simulated obstetric environment. *Anesthesiology, 106*(5), 907–915.

O'Leary, K. J., Boudreau, Y. N., Creden, A. J., Slade, M. E., & Williams, M. V. (2012). Assessment of teamwork during structured interdisciplinary rounds on medical units. *Journal of Hospital Medicine, 7*(9), 679–683. doi:10.1002/jhm.1970

Orchard, C. A., King, G. A., Khalili, H., & Bezzina, M. B. (2012). Assessment of interprofessional team collaboration scale (AITCS): Development and testing of the instrument. *Journal of Continuing Education in the Health Professions, 32*(1), 58–67. doi:10.1002/chp.21123

Peck, C., Gallucci, C., & Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education, 61*, 451–463.

Reeves, S., Perrier, L., Goldman, J., Freeth, D., & Zwarenstein, M. (2013). Interprofessional education: effects on professional practice and healthcare outcomes (update). *Cochrane Database of Systematic Reviews 2013, Issue 3. Art. No.: CD002213.* doi:10.1002/14651858.CD002213.pub3

Rethans, J. J., Van Leeuwen, Y., Drop, M., Van der Vleuten, C. P. M., & Sturmans, F. (1990). Competence and performance: Two different constructs in the assessment of quality of medical care. *Family Practice, 7*, 168–174.

Reynolds, W.M., Slattery, P., & Taubman P.M. (1996). *Understanding curriculum*. New York, NY: Lang.

Ruhstaller, T., Roe, H., Thurlimann, B., & Nicoll, J. J. (2006). The multidisciplinary meeting: An indispensable aid to communication between different specialities. *European journal of cancer (Oxford, England: 1990), 42*(15), 2459–2462.

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Annals of Surgery, 255*(4), 804–880.

Sandberg, J. (2000). Understanding human competence at work. An interpretive approach. *Academy of Management Journal, 43*(3), 9–25.

SCANS. (1999). Skills and tasks for Jobs: A SCANS report for 2000. Washington, DC: U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills (SCANS).

Schon D.A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey Bass.

Stern D. (1998). Practicing what we preach? An analysis of the curriculum of values in medical education. *American Journal of Medicine, 104*, 569–575.

Tattersall, M. H. N. (2006). Multidisciplinary team meetings: Where is the value? *The lancet Oncology, 7*(11), 886–888.

Shubert W.H., Schubert A.L., Thomas P., & W.M., C. (2002). *Curriculum books: The first hundred years* (2nd ed.). New York NY: Lang.

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58–76.

Walter, M., & Markus, G. (Eds.). (2003). *Transitions from education to Work in Europe: The integration of youth into EU Labour Markets*. Oxford, U.K.: Oxford University Press.

Walther, A., & Plug, W. (2006). Transitions from school to work in Europe: Destandardization and policy trends. *New Directions for Child and Adolescent Development, 2006*(113), 77–90. doi:10.1002/cd.170

Wass, V., Van der Vleuten, C., Shartzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet, 357*, 945–949.

Webb, N. M. (1980). Group process: The key to learning in groups. *New directions in the methodology of social and behavioral research, 6*, 77–87.

Webb, N. M. (1982). Student interaction and learning in small groups. *Review of Educational Research, 52*, 421–445.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

White, R. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333.

Whitehead, A. N. (1929). *The aims of education and other essays*. New York: Macmillan.

Zwarenstein, M., Atkins, J., Hammick, M., Barr, H., Koppel, I., & Reeves, S. (1999). Interprofessional education and systematic review: a new initiative in evaluation. *Journal of interprofessional Care 13*, 417–424.

Index

*Note:*Page numbers followed by *f* and *t* indicate figures and tables respectively

P.F. Wimmers and M. Mentkowski (eds.), *Assessing Competence in Professional Performance across Disciplines and Professions*, Innovation and Change in Professional Education 13, DOI 10.1007/978-3-319-30064-1