Comprehensive Assessment of Professional Competence: The Rochester Experiment

Ronald M. Epstein
Department of Family Medicine and
Rochester Center to Improve Communication in Health Care
University of Rochester School of Medicine
Rochester, New York, USA

Elaine F. Dannefer
Department of Medical Education and Curricular Affairs
University of Rochester School of Medicine
Rochester, New York, USA

Anne C. Nofziger
Department of Family Medicine and Rochester Center to Improve Communication in Health Care
University of Rochester School of Medicine,
Rochester, New York, USA

John T. Hansen
Department of Anatomy
University of Rochester School of Medicine,
Rochester, New York, USA

Stephen H. Schultz
Department of Family Medicine
University of Rochester School of Medicine,
Rochester, New York, USA

Nicholas Jospe
Department of Pediatrics
University of Rochester School of Medicine,
Rochester, New York, USA

Laura W. Connard
Department of Medical Education and Curricular Affairs
University of Rochester School of Medicine
Rochester, New York, USA

Sean C. Meldrum
Department of Family Medicine and Rochester Center to Improve Communication in Health Care
University of Rochester School of Medicine
Rochester, New York, USA

Lindsey C. Henson
Department of Medical Education and Anesthesiology
Case Western Reserve University
Cleveland, Ohio, USA

Background: *A required 2-week comprehensive assessment (CA) for 2nd-year medical students that integrates basic science, clinical skills, information management, and professionalism was implemented.*
Description: *The CA links standardized patients (SPs) with computer-based exercises, a teamwork exercise, and peer assessments; and culminates in student-generated learning plans.*
Evaluation: *Scores assigned by SPs showed acceptable interrater reliability. Factor analyses defined meaningful subscales of the peer assessment and communication rating scales. Ratings of communication skills were correlated with information gathering, patient counseling, and peer assessments; these, in turn, were strongly corre-*

We are grateful for funding through Title VII Predoctoral Training Grants in Family Medicine (Public Health Service Grant # 5 D16HP00003-02) and the Foundation for Innovations in Post-Secondary Education (Grant # B116000468). Brian Hodges, MD and Daniel Klass, MD contributed ideas and critiques that were essential to the success of the Comprehensive Assessment. Other University of Rochester faculty who were instrumental in the early development of the CA were Mary Ann Courtney, PhD; Harold Smith, PhD; Ralph Jozefowicz, MD; Jeffrey Lyness, MD; and Kathryn Markakis, MD. The evidence-based medicine components and script concordance questions were developed by Robert Holloway, MD, MPH and Katy Nesbit, MLS. Tana Grady-Weliky, MD offered helpful additions to the article. As Senior Associate Dean, and Dean at the University of Rochester, Edward Hundert, MD provided invaluable intellectual input and support for development and implementation of the CA.

Correspondence may be sent to Ronald M. Epstein, MD, University of Rochester School of Medicine and Dentistry, 885 South Avenue, Rochester, New York 14620, USA, Phone: 585–506–9484 X 205, Fax: 585–273–2245, E-mail: ronald_epstein@urmc.rochester.edu

lated with the written exercises. Students found the CA fair, with some variability in opinion of the peer and written exercises. Useful learning plans and positive curricular changes were undertaken in response to the CA results.
Conclusion: *A CA that integrates multiple domains of professional competence is feasible, useful to students, and fosters reflection and change. Preliminary data suggest that this format is reliable and valid.*

Teaching and Learning in Medicine, *16*(2), 186–196

The past decade has witnessed an expansion of the definition of professional competence and methods to assess it.[1] Current assessment methods have evolved from subjective assessments and multiple-choice tests to the use of simulations; and, more recently, tests of clinical reasoning.[2,3] In the process of a major curricular reform at the University of Rochester School of Medicine and Dentistry, we sought to synthesize these new perspectives and develop, *de novo,* an assessment system for 2nd- and 3rd-year medical students.

The design of our comprehensive assessment (CA) was guided by new ways of understanding professional competence. In a recent article, we defined *competence* as the "habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served"[1] (p. 226). Our definition is concordant with a parallel effort by the Accreditation Council on Graduate Medical Education and the American Board of Medical Specialties, who have mandated the demonstration of professional competence in six domains: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice.[4] All of these perspectives on competence depart from prior definitions in two ways. First, they reflect recent trends in education that focus on educational outcomes, not just course content. Second, these views maintain that competence, beyond the most elementary levels, is conferred not just by possessing a set of measurable "competencies," such as whether the clinician can ask open-ended questions or test for rebound tenderness; rather, it is the integration of several domains of knowledge, and the use of experience and judgment.

Although we use the word competence to describe the desirable outcome of professional education, Dreyfus[5] suggested several categories of competence that progress from rule-based novices to competent practitioners who can exercise clinical judgment in typical situations,[6] to expert practitioners characterized by the ability to act prudently in ambiguous, complex situations.[7,8] Others have distinguished *competence,* an inferred quality based on assessments in artificial situations, from *performance,* the habits of everyday practice.[9,10] Fraser and Greenhalgh[11] argued that the goal of training should not be competence, but rather, *capability*—the capacity of individuals to adapt to change, generate new knowledge, and continue to improve their performance. These philosophical frameworks suggest the need for a system to take into account the developmental progress of learners, and to assess performance and capability even at a relatively early stage of clinical training.

In this article, we present results of a new assessment system, the CA, which proposes to assess habits of competence in 2nd-year medical students participating in a new curriculum at the University of Rochester School of Medicine and Dentistry. Rather than trying to isolate the separate areas of competence, we created exercises that would demonstrate their integration. We used many tested and previously validated formats (standardized patients [SPs], multiple-choice questions), in addition to newer formats (peer assessments, teamwork exercises) where data regarding validation and reliability is lacking. In this article, we describe the CA, report data on feasibility, preliminary data on validity, student perceptions, and its effect on student learning and curricular change. A similar assessment for 3rd-year students, as well as formal psychometric research on elements of the CA, are reported separately.

Methods

The Context: Description of the Curriculum

The "double helix" curriculum integrates basic science and social science classroom exercises with clinical experience across all 4 years of medical school. The first 4-week segment of the curriculum introduces students to medical informatics and evidence-based medicine. During the first 2 years, 30% of curricular time is devoted to clinically relevant experiences, including an introductory clinical medicine course in the first semester of medical school, followed by 44 weekly ½-day ambulatory patient care sessions in a primary care office. In addition, there are weekly 3-hr subspecialty clinics and weekly 3-hr clinical integration conferences. The basic science courses are interdisciplinary, collaborative, and condensed during the first 2 years with additional "advanced basic science" blocks in the 3rd year. Approximately 30% of the classroom exercises are lecture format, and 70% are

small-group activities during all 4 years. The themes of aging, diversity, ethics and law, health economics, nutrition, and prevention are woven throughout the curriculum. There is a strong student advising system; for advisory deans are each assigned responsibility for 25% of each class of 100 students.[12]

Development of the CA

As part of the curriculum reform described earlier, a design team of 12 faculty members from basic science and clinical departments was assigned to develop the CA over a 2-year period. Initial meetings focused on educational research on the assessment of competence. We came to a shared understanding of the merits of different approaches and the degree to which the CA would be summative or formative.[1] We were given a 2-week period during March of Year 2 of the 4-year medical school curriculum. The rationale was that an intensive period of reflection and feedback could best demonstrate the integration of basic and clinical aspects of medicine, and also have a strong formative effect, more so than if the individual elements were presented separately. Individual exercises were developed by small groups of clinicians, basic scientists, social scientists, and educators; these were piloted during the 2000 to 2001 academic year. Student feedback informed modifications in the structure and content. We obtained grant funding to assist with the development process.

The 2nd-Year CA

The CA is a 2-week required, full-time experience for 2nd-year students. It encompasses the clinical, basic science, and social science courses during the first 2 years of medical school; and emphasizes how students integrate the knowledge, attitudes, and skills they have acquired to solve meaningful patient problems in common clinical contexts, and on how well they have learned how to learn. The CA focuses on acquiring and interpreting data, contextualized clinical reasoning, and interpersonal relationships rather than memorization. For example, students may be asked to evaluate a patient with undifferentiated chest pain using their knowledge of cardiovascular pathophysiology, the musculoskeletal exam, relation between psychological factors and symptom presentation, and the pharmacology of pain medications. The CA itself does not replicate the content or purpose of licensure examinations; students are given a practice board examination during the CA period, and take the actual examination several months later. Approximately 45% of the content is related to clinical skills, including decision making; 35% to basic science knowledge and application; and 20% to curricular themes such as aging, diversity, ethics and law, health economics, nutrition, and prevention. We also divided the content by clinical context: New acute problems appeared in 75% of the exercises, chronic illnesses in 63%, emergencies in 13%, and prevention in 38%; most exercises included more than one problem. Finally, we divided the exercises based on the level of assessment: The majority of effort in each exercise tested the application of knowledge and demonstration of behaviors rather than recall; we considered the peer assessments to give inferential knowledge about actual behavior in real-life settings.

Near the end of the CA, students receive a set of reports that guide the construction of an individualized learning plan (ILP), with the goal of creating a concrete plan to address areas of weakness, usually by taking electives or independent study. Advisory deans help students develop the ILP and monitor their progress in implementing their ILP.

The Components of the CA

The CA consists of nine components culminating in the ILP (Figure 1). The first is a series of eight 20-min Standardized Patient (SP) exercises, designed to simulate actual outpatient clinical situations. Several skills are tested in each station, and each encounter is videotaped for future individual and group review. The general format is to approach a patient with undifferentiated symptoms; prioritize the patient's concerns; understand the patient's perspective; perform necessary physical examination maneuvers; and, in some cases, engage in patient education and counseling. One of the SP exercises is preceded by a *pre-encounter exercise*. During these, students conduct literature searches and study materials necessary to prepare to counsel a patient facing a complex decision.

Immediately following each SP exercise, students do a 30-min *post encounter probe*—multiple-choice, short-answer, and brief essay questions

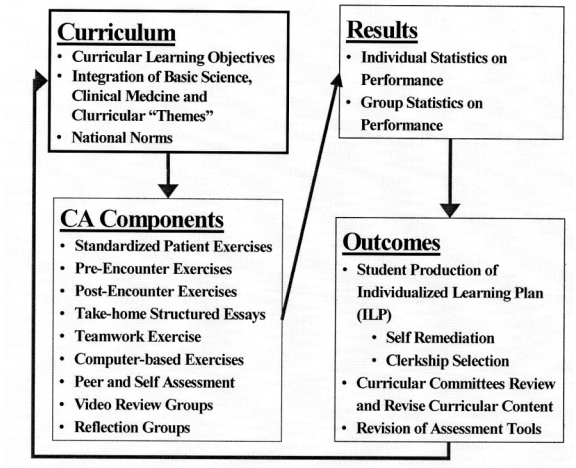

Figure 1. *Second-year comprehensive assessment, University of Rochester.*

to test students' application of principles of basic science, pathophysiology, ethics, pharmacology, or clinical decision making to the SP exercise just completed. We also use "script concordance" questions, which test students' clinical judgment by comparing their responses to the responses of a set of expert practitioners.[2,3]

Take-home exercises focus on information retrieval, interpretation, and presentation; and are directly related to each SP exercise. For example, students might be asked to do a literature search about a novel treatment, evaluate the literature to demonstrate their critical appraisal skills, or make an evidence-based recommendation for the patient whom they have seen earlier that day. One of the take-home exercises was a structured evidence-based medicine exercise, which is reported separately.[13] Other exercises included writing a chart note or a referral letter.

The *team-based human patient simulator exercise* requires a team of four students to approach a complex physiological problem using a human patient simulator.[14] It can display a wide variety of cardiovascular and respiratory derangements and respond to treatments in a way that can reproduce, in real time, what might actually occur in an intensive care unit setting. Students are directly observed by two faculty physicians who rate the process of teamwork and complex problem-solving skills. It is followed by a short-answer examination and debriefing of the team interaction.

Computer-based clinical exercises consist of computer-generated, clinically relevant exercises such as identification of dermatological lesions or microbial pathogens.

The *peer assessments* are confidential, anonymous ratings by 15 classmates, stratified by gender and ethnicity. Students are told that the purpose of the peer assessment is to reflect on aggregated subjective ratings and comments about their interpersonal habits, work styles, communication, and teamwork.[15] They are assured confidentiality and anonymity; however, some students chose to sign their comments. Because of the occasional gratuitous and destructive comments, a team of five faculty screens the student comments (with identifying data removed). Each student receives a computer-generated report that includes student and class means, standard deviations, and distribution of responses for each item. In addition, students receive raters' narrative comments. The student is the only person who receives a copy of his or her peer assessment summary report, but the student is expected to discuss it with an advisory dean and to incorporate relevant elements into his or her ILP. The *self-assessment form* is similar to the peer assessment and, thus, provides students with opportunities to compare self-ratings with those of their peers.

The final 2 days of the CA are devoted to exercises in which students review their performance on the CA and develop ILPs. Students participate in "reflection groups" in which students discuss how their values have evolved during medical training; "video review groups" in which a faculty facilitator reviews a segment of each student's videotape; and individual and small-group time for development of the ILP.

ILPs help students organize future learning by drawing attention to their strengths, weaknesses, learning styles, resources, and priorities. ILPs are "learning contracts"[16–18] that use a simple form to list (a) between three and eight specific learning objectives (such as "learn to distinguish heart sounds"), (b) methods for achieving the objectives (such as "rounding with a cardiologist at a nursing home"), (c) target completion date, and (d) methods for self-assessment and verification. Students are instructed that ILPs should be based on the results of the CA, but that they can also include elements identified during other parts of their training. Students discuss the ILP in small groups before completing them, then meet individually with their advisory dean during the ensuing 6 weeks to refine their plan. At least one elective choice for the 3rd year must be based on the ILP. The success of students' efforts are reviewed with their advisory dean later in the academic year and revisited in the 3rd-year CA.

Measures. To be able to discriminate among students, and to provide measures that were not demoralizing, we sought instruments that would generate class norms of .60 to .90, with normal distributions and standard deviations of .03 to .10. SPs completed checklists of discrete elements of the history (yes–no) and physical examination (performed correctly–performed incorrectly–not performed). History and physical examination items were derived from evidence-based criteria.[19] We limited the number of items by preferring those items with 60% to 90% correct responses on pilot testing; we tended to exclude items that greater than 95% students routinely got correct. The SPs also completed the 19-item Rochester Communication Rating Scale (RCRS, Table 1). The RCRS uses Likert scale responses to assess four domains of patient-centered communication identified as having a positive effect on health outcomes.[20–28] Analysis of SP encounters was on three levels: item (e.g., correct performance of a Murphy's sign), subcategory (e.g., physical examination skills), and category (clinical skills). Means, standard deviations, medians, and quintiles were calculated.

Team exercises were graded on a team-specific behavioral checklist by two observing physicians, who rated the team independently, then came to consensus before the debriefing session with that team. The checklist was an adaptation to 2nd-year medical students of a rating system of team skills for anesthesia crisis management cases.[29] The peer assessment measure was developed, piloted, and revised based on student and advisory dean feedback (Table 2).[15] Com-

Table 1. *Rochester Communications Rating Scale (RCRS): Items, Factor Analysis, Validation and Results From the Second-Year Comprehensive Assessment, 2002*

	M[a]	SD
RCRS Factor 1: Physician interest in patient as a person	4.8	0.3
Attended to my physical comfort during interview and physical exam (e.g., offered tissue, pulled out leg rest, warmed stethoscope, assisted me with difficult movements)	4.6	0.5
Body language and tone of voice communicated caring and concern.	5.1	0.4
Did not seem distracted.	5.1	0.3
Made an effort to understand my feelings and emotions.	4.7	0.4
Made me feel I could tell him or her anything, even something personal.	4.5	0.4
Took interest in even my smallest problems and concerns.	4.8	0.4
RCRS Factor 2: Understanding patients' experience of illness	5.0	0.2
Allowed me to tell my story in my own words.	5.1	0.2
Asked about all of my concerns early in the interview (usually by asking "anything else?").	4.9	0.3
First asked about my general concerns, then asked about specific details.	4.9	0.3
Greeted me warmly.	5.2	0.2
Let me explain my problem without interruption.	5.1	0.2
RCRS Factor 3: Attention to context	3.6	0.5
Asked about issues that affect my health, like family, culture, finances, work environment, access to care, alternative medicine, or spiritual beliefs.	3.3	0.5
Asked how the illness affects my life at home or at work.	3.7	0.5
RCRS Factor 4: Participation in care	4.8	0.3
Asked me if I had any questions.	5.1	0.3
Checked to see if I was willing and able to follow through with the treatment plan.	4.8	0.3
Clearly explained my problem and its treatment using language that I could understand.	5.0	0.3
Encouraged me to participate in treatment decisions to the extent I wished.	4.9	0.3
Summed up and made sure they understood what I said (without putting words in my mouth).	4.7	0.3
Tried to understand how I see my illness or problem ("What do you think is going on?", "What worries you the most?", "What were you hoping we would do next?").	4.5	0.4
Rochester Communication Rating Scale Overall Score	4.8	0.2

[a]All questions were scored on a 6-point scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). [b]Cronbach's $\alpha = 0.85$. [c]Cronbach's $\alpha = 0.78$. [d]Cronbach's $\alpha = 0.76$. [e]Cronbach's $\alpha = 0.81$.

puter-based multiple-choice exercises were graded automatically. Short-answer and essay exercises were graded by residents, fellows, and 4th-year students using a prepared answer guide that indicated key points to be emphasized. Literature searches were graded according to criteria that rewarded basic competency,[13] with care not to penalize proficient students who took effective shortcuts.

To evaluate the CA in 2001, 49 of the 98 students completed a 50-item, online assessment form in the 2 weeks following the CA. Although they are similar to the 2002 results, because of potential sampling bias, we chose not to report these. In 2002, all 95 students completed a 61-item, online assessment at the end of the CA, with multiple opportunities for comments. Each year, students participated in focus groups; each group was devoted to a different aspect of the CA.

Results

We report results of the 2002 CA.

SP Rating Scales

SP interrater reliability was determined by having a second SP review the videotapes. Interrater reliability (intraclass correlation) was .80, .84, and .82 for history, physical examination, and counseling, respectively; and .59 for the RCRS. The RCRS showed item means from .45 to .83, with standard deviations of .03 to .10. Crohbach's alpha was .91, indicating high internal consistency of the scale. Table 1 displays the items listed by factor, internal consistency of the subscales, and student results.

Information gathering and physical examination scales were normally distributed with means of .75 and .71, respectively. Generic questions about onset, severity, and location of the symptom received significantly higher scores than disease-specific inquiries that would require higher order diagnostic hypothesis testing during the interview (e.g., family history, associated symptoms, presence of fever, prior episodes), $t = 23.96$, $p < .0001$. Similarly, specific physical examination maneuvers (Murphy's sign in right upper quadrant pain, ankle dorsiflexion strength in sciatica) were performed less frequently than generic maneuvers (palpation, auscultation), $t = 17.77$, $p < .0001$.

Other Exercises

The take-home exercises showed that students' ability to apply knowledge of basic mechanisms of disease to the conditions portrayed by the SPs was case de-

Table 2. *Peer Assessment—Second Year Comprehensive Assessment, 2002*

Low–Unsatisfactory		High–Exceptional
Peer Assessment Factor 1: Work habits		
Consistently seems unprepared for sessions.	1 2 3 4 5 na	Consistently well-prepared for sessions; presents extra material; supports statements with appropriate references.
Overlooks important data and fails to identify or solve problems correctly.	1 2 3 4 5 na	Identifies and solves problems using intelligent interpretation of data.
Unable to explain clearly his/her reasoning process with regard to solving a problem, basic mechanisms, concepts, etc.	1 2 3 4 5 na	Able to explain clearly his/her reasoning process with regard to solving a problem, basic mechanisms, concepts, etc.
Lacks initiative or leadership qualities.	1 2 3 4 5 na	Takes initiative and provides leadership.
Only assumes responsibility when forced to or stimulated for personal reasons; fails to follow through consistently.	1 2 3 4 5 na	Seeks appropriate responsibility. Consistently identifies tasks and completes them efficiently and thoroughly.
Dependent upon others for direction with regard to his/her learning agenda.	1 2 3 4 5 na	Directs own learning agenda; able to think and work independently.
Peer Assessment Factor 2: Interpersonal sensitivity		
Lacks appropriate respect, compassion and empathy.	1 2 3 4 5 na	Always demonstrates respect, compassion and empathy.
Displays insensitivity and lack of understanding for others' views.	1 2 3 4 5 na	Seeks to understand others' views.
Doesn't share information or resources; impatient when others are slow to learn; hinders group process; tends to dominate the group.	1 2 3 4 5 na	Shares information or resources; truly helps others learn; contributes to the group process; able to defer to the group's needs.
Does not seek feedback; defensive or fails to respond to feedback.	1 2 3 4 5 na	Asks classmates and professors for feedback and then puts suggestions to good use.
Pleases superiors while undermining peers; untrustworthy.	1 2 3 4 5 na	Presents him/herself consistently to superiors and peers; trustworthy.
Hides his/her own mistakes; deceptive.	1 2 3 4 5 na	Admits and corrects his/her own mistakes; truthful.
Additional Items		
Dress and appearance often inappropriate for the situation.	1 2 3 4 5 na	Dress and appearance always appropriate for the situation.
Behavior is frequently inappropriate.	1 2 3 4 5 na	Behavior is always appropriate.
I have concerns for his/her future patients.	1 2 3 4 5 na	I would refer my own family or patients to this future physician or ask this person to be my own doctor.

pendent and inconsistent. Team performance on the human patient simulator exercise achieved a mean score of .72 (SD = .03), with a normal distribution. Performance on the comprehensive basic science examination showed a normal distribution (M = .64, SD = .06).

The peer assessment showed a mean score of 4.2 (range = 1–5, SD = .3) based on an average of 14.3 reviews for each student (range = 7–16, Mdn = 15).[15] We have previously reported that factor analysis of the peer assessments revealed two factors ("work habits" and "interpersonal sensitivity") and a single item ("I would refer friends or patients to this future physician…").[15]

ILPs

Although the adequacy of ILPs can only be judged in the context of each student's overall performance and his or her ability to carry through with stated goals, qualitative comments from the advisory deans indicated that all students have completed the ILPs; and that they include goals derived from the SP, written, peer assessment, and video review exercises. Advisory deans reported that the depth of thoughtfulness of learning objectives was mirrored by the degree to which students used the peer assessments in their ILPs. Student focus groups reported that the ILPs were valuable, and suggested some logistical changes to their earlier introduction to the concept.

Correlations Between Elements of the CA

Correlation analyses, regression analyses, and factor analyses were performed. The factor structures of the RCRS is shown in Table 1, and corresponds to current understandings of patient-centered care.[20–28] A complete correlation table of all the elements of the CA is in Table 3. We were most interested in correlations among the newer scales. The RCRS showed moderate to strong correlations with information gathering (r = .47, $p < .001$) and patient counseling (r = .50, $p < .001$) items as rated by the SP, and moderate correlations with the peer assessment (r = .36, $p < .001$). The RCRS correlated moderately with the post-encounter and take-home written exercises (r = .31, p = .002), but this effect disappeared when multiple regression analyses were performed. The peer assessment was strongly correlated with the post-encounter and take-home written exercises (r = .49, $p < .001$); this effect was entirely due to the work habits factor. The interpersonal sensitivity factor did not correlate with any of the other measures, indicating that it measures a separate domain. We are awaiting

Table 3. *Pearson Correlations Among Components of the Comprehensive Assessment Examination*

	Information Gathering	Physical Exam	Education & Counseling	RCRS	PEP–Take-Home Essays	Simulator–Teamwork	Computer Exercises	Peer Assessment
Information gathering	1.0	0.05	0.50[c]	0.47[c]	0.18	0.00	0.17	0.15
Physical examination		1.0	0.00	0.11	0.24[a]	0.15	0.15	0.22[a]
Patient education and counseling			1	0.50[c]	0.29[b]	0.08	0.23[a]	0.20[a]
RCRS				1.0	0.31[b]	0.00	0.08	0.33[c]
PEP/take-home essays					1.0	0.13	0.45[c]	0.43[c]
Simulator–teamwork						1.0	0.16	0.15
Computer exercises							1.0	0.28[b]
Peer assessment								1.0

Note: $N = 97$ students. RCRS = Rochester Communication Rating Scale; PEP= post-encounter probes.
[a]Prob > |r| under H0: Rho = 0 is less than 0.05. [b]Prob > |r| under H0: Rho = 0 is less than 0.01. [c]Prob > |r| under H0: Rho = 0 is less than 0.001.

results from subsequent years to report further psychometric characteristics of the measures.

Student Evaluation of the CA. The CA was rated at a level equivalent to the highest rated courses in the first 2 years of medical school (Table 4). The SP exercises with video reviews were realistic and allowed students to reflect on their strengths. The simulator exercise was considered a good test of teamwork and clinical reasoning skills. The post-encounter and essay exercises were perceived to be directly related to the SP encounter, addressed issues that the students viewed as important, and were at the right level of difficulty. Focus group comments called for even stronger links between basic science questions and clinical exercises, time limits on essay questions, and improvement of some of the computer images. Students complained that the focus on evidence-based medicine was excessive prior to the CA;[13] a possible reason why they rated the components favorably but did not find the exercise valuable. The majority of students reported that the peer assessment exercise was a helpful way to receive feedback. Student comments informed modifications in the structure and content of the CA during the pilot phases, from 2001 to 2002, and beyond. These included expanding the post-encounter probes from 5 to 30 min, a greater focus on clinical reasoning, changing the written essays from in class to take home, changing the standards for evaluating literature searches, incorporating evidence-based medicine into other exercises, simplifying the presentation of the results to the students, and instituting a structured oral examination (in 2003).

Table 4. *Second-Year Student Ratings of the 2002 Comprehensive Assessment*

	Strongly Disagree or Disagree (1 or 2)	Mixed Feelings (3)	Agree or Strongly Agree (4 or 5)	M	SD
"The Comprehensive Assessment was a fair way to assess my overall knowledge, skills, and attitudes."	6	27	63	3.65	0.68
SP encounters					
Portrayals were realistic	1	22	73	3.97	0.70
Presenting problems similar to those encountered in primary care clerkship	13	28	55	3.57	0.93
Pre-encounter exercise					
Medline prepared me to counsel patient	2	14	80	4.11	0.77
My literature search would allow me to make better clinical decisions in this case.	4	6	86	4.20	0.78
Reinforced value of good EBM skills	4	25	65	3.82	0.84
Post-encounter exercises					
Directly related to the SP encounter	3	16	77	3.99	0.76
Addressed important issue	5	26	65	3.78	0.77
EBM take-home exercise					
Exercise directly related to SP case	11	31	54	3.63	0.94
Addressed important issues	16	36	44	3.27	1.00
Good test of Medline searching	17	22	57	3.47	0.98
Good test of critical appraisal skills	20	29	47	3.23	0.97
Reinforced value of good EBM skills	24	26	45	3.17	1.12
Take-home exercises					
Exercise directly related to SP encounters	1	18	75	3.94	0.67
Exercises required applied basic science	6	32	58	3.57	0.83
Addressed important issues	16	25	55	3.48	0.96
Exercises identified learning needs	37	28	31	2.88	1.07
Simulator exercises					
Tested my clinical reasoning skills	8	16	72	4.01	0.92
Good test of my ability to work on a team	5	13	78	4.06	0.92
"How valuable of a learning experience was each component of the Comprehensive Assessment?"					
SP encounters	1	8	87	4.25	0.65
Individual video review session	2	20	73	4.11	0.84
Simulator exercise	9	19	68	3.90	0.99
Small group video review session	7	28	61	3.75	0.88
Individualized learning plan	6	36	54	3.65	0.88
Post-encounter exercises	6	30	59	3.61	0.78
Pre-encounter search	8	40	48	3.48	0.82
Peer/self-assessment exercise	19	27	49	3.42	1.20
Other take-home exercises	33	46	16	2.75	0.86
EBM take-home exercise	53	33	9	2.31	0.97

Note: SP = standardized patient; EBM = evidence-based medicine.

Effect of the CA on the Medical School Curriculum

Several important curricular changes were undertaken in response to the CA results. Faculty development courses for preceptors and a monthly "master clinician seminar" were created to further reinforce clinical reasoning skills. Expectations of 1st- and 2nd-year students in clinical settings were clarified. Greater integration between basic science and clinical courses was promoted by discussion of the results of the CA, leading to further development of the collaborative "integration conferences;" shared examinations between basic science and clinical courses; and structuring of 2-week basic science blocks at the end of each 3rd-year clinical rotation, to further reinforce connections between basic knowledge and clinical practice.

Discussion

We implemented an ambitious CA format that was linked to an integrated curriculum. The CA format was based on current educational research and made efficient use of new assessment technologies. The CA was well-accepted by students; tested a broad range of content domains; successfully focused on the integration of knowledge, skills, and professionalism; and provided data on student performance that effectively informed future learning plans and curricular design. The domains evaluated represent enduring qualities of excellent physicians, rather than testing recall of knowledge that is likely to become out of date. The new subjective rating scales developed for the CA (the RCRS and the peer assessment) added useful information not contained in other assessment formats.

In designing the CA, we took into account four principles of assessment of professional competence.[1] First, we developed a wide variety of clinical contexts to address the concern that competence is context dependent. Second, we tested not only knowledge but also the ability to engage in ongoing learning. Third, we sought to respond to the debate about validity and reliability of tests of competence.[30–32] To date, the most reliable tests (multiple-choice examinations) encompass only a small portion of the skills, attitudes, and knowledge relevant to professional practice; and performance on such examinations does not necessarily predict the quality of future clinical care. Similarly, subjective evaluations in real-life or life-like settings achieve face and construct validity, but may take multiple observations in different assessment formats to achieve acceptable reliability. Therefore, whenever possible, we chose to use validated measures that correlate with patient outcomes, and to triangulate observations about attributes, such as professionalism, that are difficult to measure. Fourth, assessments at a clerkship or institutional level are different from assessments for licensure at the state or national level, in that licensure examinations protect the public primarily by assuring a minimum level of knowledge. In contrast, institution- or course-level assessments such as the CA provide information to guide promotion decisions, foster future learning, and inform curricular revisions.

Combined methods may be one way of achieving the goals of providing reliable and valid assessments at both the national and local level. Many U.S. specialty boards still include an oral examination in addition to multiple-choice tests.[33] Several Canadian boards use a combined multiple-choice–key features written examination linked with a SP examination, and have achieved better reliability and validity than either element alone.[34–36] The US Medical Licensing Examination has adopted a similar SP assessment format, to be implemented in 2004. As a result, more medical schools will likely adopt combined assessment programs. For example, the University of New Mexico and others have developed periodic assessments using multiple formats for medical students.[37] Formal assessment methods are less well developed at the residency level, but are now required by the Accreditation Council on Graduate Medical Education.[4]

Assessments that claim objectivity on the basis of high reliability may create a false sense of validity.[30,31] It has been suggested that multimethod assessments may collectively achieve equivalent reliability to single-format examinations, as well as provide a more valid and complete picture of overall competence. We have operationalized this approach to assessment by combining subjective assessments (including peer assessments, SP ratings, and written essays) with SP checklists, multiple-choice questions, and script concordance tests. Combined assessments also set a more holistic standard of professional competency that moves beyond domain competency, and may reflect the qualities characteristic of professional expertise. Furthermore, multiple subjective SP and peer assessments may give insight into habits of mind that confer not only competence but capability to deal with new and complex problems.

In our experience, the linking of SP cases with post-encounter exercises reinforced and further informed the curricular goals of integrating clinical and basic knowledge. Despite a highly integrated curriculum and temporal linking of tests of clinical and basic knowledge, students mentally linked the written exercises with the clinical exercises only some of the time. Substantial descriptive research indicates that expert clinicians commonly use pattern recognition and iterative hypothesis testing, but that these processes function well only when basic knowledge is adequately structured.

Assessments that test the application of basic knowledge to clinical care invite critical exploration of the nature of the so-called basic sciences, and in what sense they relate to expert clinical care. Parts of basic science knowledge may comprise, as Polanyi[43] suggested, a body of tacit knowledge that informs clinical decisions, but is rarely brought to a level of explicit discourse. Taking this view, however, many of the principles used by expert clinicians may not be reinforced by basic science courses, which tend to emphasize solving of idealized problems with clear-cut solutions, rather than basic scientific inquiry that requires creative thinking and perseverance. These courses may not be involving students in decisions that require judgment to navigate the often ambiguous, emotionally laden controversies of daily clinical practice. We would invite further discussion of what is a basic science in current medical practice, and how medical education might redefine basic science to reflect more accurately those principles that actually inform clinical actions.

Students and advisory deans noted the advantages of combining the assessment of interpersonal attributes and some elements of professionalism with cognitive and technical ability. The peer assessment, the video review of the SP cases, and the SP assessments of communication were taken very seriously by students because they were linked to solving meaningful clinical problems in realistic contexts. Assessment of these attributes has been difficult, and our approach has overcome some of the concerns about applicability, feasibility, and impact. However, care must be taken in how peer data are used. Students commented that they will only be honest and value the feedback that they receive if they have some control over the data, and if they feel that it is presented in an environment of trustworthiness and respect.

Future directions include incorporation of live faculty examiners into a CA. Several schools[44] use faculty oral examiners for 5 to 7 min, post-encounter probes to test clinical reasoning. Our approach, using written exercises, had the advantage of allowing more consistency at the expense of losing the potentially valuable subjective comments of faculty. A combined approach, using written and oral post-encounter assessments, may provide the most useful information.

Given that our curricular reform included a strong advisory system, we were not surprised by feedback that there were few, if any, decisions about promotion from the 2nd to the 3rd year of medical school that would have changed as a result of the CA, beyond the information already available. This finding supports the continued use of the CA to provide further characterization of marginal performance and to guide students' further learning, versus identifying, *de novo,* students with severe academic difficulty. This conclusion may not be generalizable to medical schools without student advising systems and small-group learning formats in which student performance can be closely monitored.

Limitations

The effectiveness of our approach will be validated by the effect that it has on outcomes in real clinical settings—the health of the future patients for whom our students will care. Clearly, this is a difficult standard to apply when assessing students in the 2nd year of medical school. However, we believe that use of rating scales of communication and physical examination that have been correlated with patient outcomes in real clinical settings may provide reasonable intermediate outcomes, as might script concordance questions and other tests of clinical reasoning.[2,3] Overall test–retest reliability for a multidimensional examination such as ours requires a large sample size, which will only be achieved after at least 10 administrations, during which time modifications can and should occur; this would be remedied if there were several schools with similar curricula that could administer the same CA simultaneously. Standard setting for SP exercises is a complex matter, and setting standards for exercises that propose to demonstrate links between different domains of knowledge is even more difficult. With the expansion in use of multidimensional testing, this will be an important field of inquiry. Implementation of a new curriculum and assessment system is resource intensive, and commitment of the senior leadership is instrumental to its success.

Conclusions

We have developed and demonstrated a means to assess medical students across numerous domains using a comprehensive model of competence that integrates basic science knowledge, clinical skills, teamwork, communication, and professionalism. This format characterized students' ability to acquire and use knowledge, apply basic science principles to clinical settings, and make clinical judgments in greater detail than simple multiple-choice and SP examinations. Peer assessments and teamwork exercises added dimensions not measured by other means of assessment at the 2nd-year student level. Students' ILPs influenced their future learning choices. The results of the CA influenced the structure and content of the medical school's curriculum. Success of CAs such as ours depends on visible involvement of faculty in creating and administrating the assessment, a good student advising system, and creating an atmosphere of trust between the leadership and the students. Dissemination of the CA should take into account other institutions' unique curricula and advising systems.

References

1. Epstein RM, Hundert EM. Defining and assessing professional competence. *Journal of the American Medical Association* 2002;287:226–35.
2. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van D. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: An experimental study on the script concordance test. *Medical Education* 2001;35:430–6.
3. Charlin B, Roy L, Brailovsky C, Goulet F, Van D. The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine* 2000;12:189–95.
4. *Accredation Council for Graduate Medical Education. ACGME Outcomes Project.* Retrieved March 1, 2000, from http://www.acgme.org
5. Dreyfus HL. *On the Internet (thinking in action).* New York: Routledge, 2001.
6. Flyvbjerg B. *Making social science matter.* Cambridge, England: Cambridge University Press, 2001.
7. Schon DA. *Educating the reflective practitioner.* San Francisco: Jossey-Bass, 1987.
8. Schon DA. *The reflective practitioner.* New York: Basic Books, 1983.
9. Rethans JJ, Sturmans F, Drop R, van der Vleuten A. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *British Journal of General Practice* 1991;41:97–9.
10. Ram P, van der Vleuten CPM, Rethans JJ, Grol R, Aretz K. Assessment of practicing family physicians: Comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. *Academic Medicine* 1999;74:62–9.
11. Fraser SW, Greenhalgh T. Coping with complexity: Educating for capability. *BMJ* 2001;323:799–803.
12. Grady-Weliky T. *Student advising at the University of Rochester.* AAMC Northeast Group on Student Affairs, Toronto, 2002.
13. Holloway R, Nesbit K, Bordley D, Noyes K. Teaching and evaluating first and second year medical students' practice of evidence-based medicine. *Medical Education* in press.
14. Human Patient Simulator, Model C. Sarasota, FL: Medical Education Technology, Inc., 2002.
15. Dannefer EF. *Peer and self assessment of professionalism for medical students.* Ottawa Conference on Medical Education and Assessment, Ottawa, Canada, July 15, 2002.
16. Knowles M. *Self-directed learning: A guide for learners and teachers.* New York: Cambridge Book Co., 1975.
17. Knowles MS. *The modern practice of adult education: From pedegogy to andragogy.* New York: The Adult Education Company, 1980.
18. Westburg J, Jason H. *Collaborative clinical education: The foundation of effective patient care.* New York: Springer, 1993.
19. Sackett DL. *Evidence-based medicine: How to practice and teach EBM.* New York: Churchill, 1997.
20. Mead N, Bower P. Patient-centredness: A conceptual framework and review of the empirical literature. *Social Sciecne and Medicine* 2000;51:1087–110.
21. Stewart M, Brown JB, Weston WW, McWhinney IR, McWilliam CL, Freeman TR. *Patient-centered medicine: Transforming the clinical method.* Thousand Oaks, CA: Sage, 1995.
22. Makoul G. Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Academic Medicine* 2001;76:390–3.
23. Williams GC, Freedman Z, Deci EL. Promoting motivation for diabetics' self-regulation of HgbAlc. *Diabetes* 1997;45:13A.
24. Greenfield S, Kaplan SH, Ware JE, Jr, Yano EM, Frank HJL. Patients' participation in medical care: Effects on blood sugar control and quality of life in diabetes. *Journal of General Internal Medicine* 1995;3:448–57.
25. Kaplan SH, Greenfield S, Ware JE, Jr. Assessing the effects of physician–patient interactions on the outcomes of chronic disease. *Medical Care* 1989;27:S110–27.
26. Safran DG, Taira DA, Rogers WH, Kosinski M, Ware JE, Tarlov AR. Linking primary care performance to outcomes of care. *Journal of Family Practice* 1998;47:213–20.
27. Safran DG, Kosinski M, Tarlov AR, et al. The Primary Care Assessment Survey: Tests of data quality and measurement performance. *Medical Care* 1998;36:728–39.
28. Safran DG, Montgomery JE, Chang H, Murphy J, Rogers WH. Switching doctors: Predictors of voluntary disenrollment from a primary physician's practice. *Journal of Family Practice* 2001;50:130–6.
29. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology* 1998;89:8–18.
30. van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991;25:110–8.
31. Norman GR, van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education* 1991;25:119–26.
32. van der Vleuten, CPM. Validity of final examinations in undergraduate medical training. *BMJ* 2000;321:1217–9.
33. Mancall EL, Bashook PG. *Assessing clinical reasoning: The oral examination and alternative methods.* Evanston, IL: American Board of Medical Specialties, 2002.
34. Tamblyn R, Abrahamowicz M, Brailovsky C, et al. Association between licensing examination scores and resource use and quality of care in primary care practice. *Journal of the American Medical Association* 1998;280:989–96.
35. Dauphinee WD. Assessing clinical performance. Where do we stand and what might we expect? *Journal of the American Medical Association* 1995;274:741–743.
36. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. *Journal of the American Medical Association* 2002;288:3019–26.
37. Obenshain SS. *Student Progress Assessment Manual.* Albuquerque: University of New Mexico, 1998.
38. Epstein RM. Mindful practice. *Journal of the American Medical Association* 1999;282:833–9.
39. Gruppen LD, Frohna AZ. Clinical reasoning. In GR Norman, CPM van der Vleuten, DI Newble (Eds.), *International handbook of research in medical education, Part 1* (pp. 205–230). Dordrecht, The Netherlands: Kluwer, 2002.
40. Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine* 1994;69:883–5.
41. Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: Categories and prototypes. *Medical Education* 1984;18:406–16.
42. Friedman MH, Connell KJ, Olthoff AJ, Sinacore JM, Bordage G. Medical student errors in making a diagnosis. *Academic Medicine* 1998;73:S19–21.
43. Polanyi M. *Knowing and being, the logic of tacit inference. Knowing and being: Essays by Michael Polanyi.* Chicago, IL: University of Chicago Press, 1969.
44. Hodges, B. *Assessment across basic science and clinical skills: Using OSCEs with post-encounter basic science probes.* Toronto: University of Toronto Press, 2001.

Received 20 March 2003
Final revision received 18 August 2003