

BEST PRACTICES FOR ASSESSING COMPETENCE AND PERFORMANCE OF THE BEHAVIORAL HEALTH WORKFORCE

Philip G. Bashook

ABSTRACT: The need for mechanisms to assess the competence and performance of the behavioral health workforce has received increasing attention. This article reviews strategies used in general medicine and other disciplines for assessing trainees and practitioners. The possibilities and limitations of various approaches are reviewed, and the implications for behavioral health are addressed. A conceptual model of competence is presented, and practical applications of this model are reviewed. Finally, guidelines are proposed for building competency assessment protocols for behavioral health.

KEY WORDS: assessment; behavioral health; competencies; workforce.

Assessment of health care providers' competencies occurs throughout the continuum of training and practice. Patients and clients, clinical experts, supervisors, and other health care providers informally evaluate these individuals every day. The expected competence of behavioral health care providers can be summarized in the phrase: *he/she should know his/her own limits of expertise, and should know what to do when those limits are reached.* Articulation of defined competencies by the Annapolis Coalition (Hoge, Tondora, & Marrelli, in press) translates knowing "one's own limits of expertise" as knowledge of the science of behavioral health care and how to use that knowledge in a caring and appropriate manner. One should also keep in mind that assessment of competence before entry into practice is quite different from assessment of performance in practice.

Philip G. Bashook, Ed.D., is a Research Assistant Professor at the University of Illinois at the Chicago College of Medicine.

This work was supported in part by Contract No. 03M00013801D from the Substance Abuse and Mental Health Services Administration.

Address for correspondence: Philip G. Bashook, Ed.D., Department of Medical Education, University of Illinois at Chicago, 808 South Wood Street (MC591), Chicago, IL 60612. E-mail: pgb@uic.edu.

A general schema has been proposed to assess competence of physicians and other health care practitioners (Newble et al., 1994). Using this schema assessment of competence in the behavioral health workforce should begin by defining how the assessments will be used and what assessment results will be needed. Keeping this bigger picture in mind, an assessment plan might unfold by addressing each of these questions: who is to be assessed? What should be in the blueprint of competencies to be assessed along career paths (during training, pre-practice for certification/licensure, and during practice/employment)? What combination of assessment methods can provide the best measures for each of the competencies to be evaluated (Norman, Swanson, & Case, 1996), given the available resources and the intended uses of the assessments? The paper is organized into sections that follow this approach to the assessment planning process. It concludes with recommended best practices for assessment of competencies illustrated by examples for selected members of the behavioral health workforce.

THE BEHAVIORAL HEALTH WORKFORCE

The assessment challenge is to develop and use valid and reliable assessment methods that measure the competencies relevant to the setting and roles where each member of the behavioral health workforce functions. It is impractical in this paper to recommend an assessment plan for the more than 20 different types of behavioral health specialty disciplines, not to mention customizing the applications to the hundreds of settings where they practice. The elements of the behavioral health workforce have been described in previous work of the Annapolis Coalition and the Institute of Medicine (Hoge & Morris, 2002, 2004; Morris, Goplerud, & Hoge, 2004). This article reflects the broad breakdown of the workforce into those with graduate training, those with baccalaureate training, frontline providers, and consumers and families. Within these broad categories of course, extensive variation exists among the types of licensure and certification standards along the dimensions of educational level, credentialing authority, and state regulation. At present, very few formal structures exist for credentialing consumers and family members, who are increasingly acknowledged as significant parts of the workforce. The most significant exception to this observation is the emergence of peer support specialists. The peer support specialists are newly defined members of the behavioral health workforce who are current and/or former mental health consumers. They are trained to fulfill key roles in advocacy and consumer support of Medicaid-funded mental health

services. The certified peer specialists, originally only in the state of Georgia, complete competency-based training modules, and written and oral examinations (Sabin & Daniels, 2003). Since assessment principles described here for non-degreed staff most closely apply to this group of behavioral health providers, these individuals will not be discussed further here (see the website of the Georgia Certified Peer Specialist, 2004 Project for more details at <http://www.gacps.org>).

Assessment Technology

When considering which assessment technology to use, a significant challenge is judging the quality of the possible assessment methods. The goal is to generate as many quality measurements as possible about a trainee or practitioner across as many examples as possible of the trainee/practitioners knowledge, skills, abilities, and practice performances. Assessments for high-stakes decisions like graduation, certification, and licensure, for example, must be credible. This requires proof and documentation of the reliability and validity of the assessment results. The priority for assessments during training is to select the most feasible methods (i.e., the least expensive in terms of direct costs and labor) to obtain useful data for giving constructive feedback to trainees, and for making decisions about continuing or advancing the trainee in the program.

It is helpful to know the assessment jargon when weighing the value of one assessment method over another. Commonly used concepts for judging assessment methods are the psychometric requirements of reliability and validity, feasibility, and credibility. Each concept will be discussed briefly, followed by descriptions of commonly used assessment methods and the competencies each can best measure.

Psychometric Requirements. These are the estimates of the reliability and validity of an assessment method for a specific purpose. When measuring the competencies of an individual during training or in practice, the goal is for each assessment to be an accurate measure of the person's knowledge, skills, abilities, or performance. Accuracy means that the scores from the assessment are reliable and a valid measure of that person's performance. It is important to recognize that validity does not mean a method is valid per se, but refers to the validity of what the score means when used for a specific purpose with a specific group of people. Experts in psychometrics have developed statistical tests and procedures for calculating reliability estimates. They have also devised procedures for summarizing and interpreting an accumulation of studies necessary to

establish the validity of scores derived from an assessment method (Joint Committee on Testing Practices, 1999; Linn, 1994).

The research evidence suggests these instructor-made tests are rarely reliable, and may not cover the content adequately.

Reliability. Technically, reliability means an estimate of the measurement error for an assessment score (Brennan, 2001; Joint Committee on Testing Practices, 1999). The most useful reliability statistic in assessment is a calculation of the measurement error if the same assessment were repeated under similar conditions. This estimate of measurement error is called score reproducibility (Lunz, Stahl, & Wright, 1994). A highly reliable test of knowledge, for example in a standardized test format, would have a very low error rate and be expressed as having a reliability of 0.90 or greater (i.e., good score reproducibility). The reliability scale uses 0 as unreliable and 1.0 as perfect reliability. In performance assessments and assessments of skills for high-stakes decisions, acceptable reliabilities are above 0.85. Explanations for estimating reliability of written examinations can be found in Case and Swanson (2003). For performance assessments, see Lunz (1995) or Swanson, Norman, and Linn (1995).

In complex assessments like simulations, or when combining multiple assessment methods, it is necessary to separate out the estimated reliability of the score for each person from variations due to the method used, the difficulty of the clinical cases or situations presented, the severity or easy grading by assessors/raters, and different administrations of the assessment over time and location. These variables are referred to as facets when calculating reliability with the Rasch statistical model (Andrich, 1988; Lunz et al., 1994) or components when using generalizability theory (Shavelson & Webb, 1991).

Validity. The concept of validity refers to the accumulated evidence about how well an assessment of competencies measures what it is intended to measure (Joint Committee on Testing Practices, 1999). Validity is not a single statistical calculation, but rather a construct combining statistics, observations, and logical arguments to explain the quality of the validity evidence. In psychometric terms, validity refers to the consistency of scores on an assessment with a preconceived “psychological construct” that defines a person’s abilities or explains performance in practice. In the modern concept of validity, even statistical estimates of reliability are subsumed under construct validity, because reliability influences judgments about the veracity of assessment scores.

Content validity refers to selecting the appropriate range of topics and situations for the assessment. Content validity usually involves creating a blueprint for an examination or assessment and determining that the administered assessment items match the distribution of content defined in the blueprint. In performance assessments, content validity is established by experts selecting the situations or client cases to be used in an assessment, and confirming that the sample of cases is representative of the practice (LaDuca, 1994). Evidence for concurrent validity compares performance by the same people on one assessment (e.g., a simulated case problem) with a well-established score from another assessment (e.g., ratings from training supervisors), both administered contemporaneously as much as possible. A predictive validity study about simulated client cases, for example, might establish that a measurement of a person's abilities managing simulated client cases while in training has a high correlation with performance in actual practice.

Feasibility. Feasibility can be divided into the theoretical and practical problems of design, development, and production of an assessment method, as well as the administration, data analysis, reporting, and ongoing revisions and use of the method. In nearly all situations, feasibility becomes a question of available money, expertise, opportunity, resources, and time. The most efficient approach is to borrow a proven existing method, make minor changes to adapt it for use in the new setting, and hope that the method is as valid for the new setting and the different type of health provider as it was previously. This is the least costly approach, but leaves in question the validity of the results. There is extensive literature describing the transportability of assessment methods, which pivots on one question: will doing the assessments in a new setting or with different stimulus cases/items or raters still provide reproducible and valid measures of competencies the assessment was intended to measure? (Joint Committee on Testing Practices, 1999; Linn, 1994).

Practical concerns with using any assessment method, as noted above, are the time, expertise, and resources needed to use it properly and get useful results. Most clinical settings lack one or more of these. Training settings can often customize survey or rating forms by making minor changes to existing ones. This is quite easy and can be done at minimal cost. Creating custom forms should be sufficient to document a supervisor's ratings of trainees and give trainees feedback, but may not be credible or defensible for pass/fail decisions without additional corroborative evidence of reliability and validity.

In contrast, when resources are more plentiful, as with certifying boards, it is possible to produce a full battery of methods and even have a pool of test questions that can be used year to year. Practical concerns

are cost and sustaining the quality of the assessment method to assure credible results. A complete written examination for board certification (150 high-quality test questions per half-day exam) typically takes 12–18 months for initial planning to administration. The average cost is between \$1000 and \$1500 per item for development alone (\$225,000 per test), excluding test administration and the time of voluntary experts writing test questions. A practical examination like case-based orals takes less time because fewer cases are needed, but costs slightly more, since administration of the exam requires live expert examiners (\$1500 per case, or \$500–\$1000 per candidate). Budgeting for either assessment method needs to include experts meeting to construct and review items, consultants or staff with test construction expertise, editing and revising questions, pilot testing questions, and statistical analysis to document reliability and validity, obtain statistics about the quality of each test question, and administer the assessment to candidates (Browning, Bugbee, & Mullins, 1996).

Another practical matter is administering the assessment. Written exams, for example, are shifting from paper-and-pencil, to computer-based or web-enabled delivery of exams (Mancall, Bashook, & Dockery, 1996). Computers can vividly and accurately display pictures, video clips of clients, and actual clinical findings, allowing the user to zoom in on images, repeat video clips, and move easily from question to question. There are thousands of commercially run computer testing centers in all large cities and many smaller ones (e.g., <http://www.prometric.com>, <http://www.vue.com>). For-profit and nonprofit vendors also provide exam development expertise, candidate scheduling and registration, and verification of candidates during administration. Feedback from users reflects greater satisfaction with computer-delivered tests than paper-and-pencil administrations for high-stakes tests, and they appreciate the reduced time and cost, and added convenience of local travel to test sites. On the other hand, the costs are high for administration. A half-day to one-day exam can cost over \$80 per candidate seat at a commercial testing site. Clearly, this mode of test delivery is potentially feasible for large-scale testing by certifying or licensure boards. The candidates pay the testing cost through certification fees. In contrast, paper-and-pencil test delivery is most common during training.

Successful simulations force the trainee or practitioner to sort through a wide variety of options to clarify the important clinical problems and challenges.

Credibility. Credibility refers to the veracity of assessment results from the perspective of those who will use the results (e.g., the behavioral health community, colleagues in the same discipline, the public, govern-

ment regulatory agencies, and clients). A straightforward rule of thumb for judging credibility is deciding if the assessment results are a good measure of whether the person “knows their own limits and what to do when those limits are reached.” Credibility indicates how well the assessment results are supported by affirmative answers to the following questions:

- Are the content and competencies being assessed appropriate for the providers’ expected roles and responsibilities?
- What is the appropriate use of the assessment results? Training feedback? Training promotion? Employment? Certification? Licensure? Practice privileges?
- Was appropriate scientific rigor used in the design and execution of the assessment methods and the assessment process?
- Are the assessment methods appropriate for the type of provider?
- Were any adjustments made to accommodate the providers’ disabilities?
- Is the assessment fair to all those who take it?
- Are the raw findings in the assessment results kept confidential, as appropriate?

Assessment Methods

The commonly used assessment methodology can be classified into four categories according to what each is intended to measure. Table 1 describes each method and recommended uses for assessing competencies of behavioral health providers. Some of these descriptions build on the ACGME Toolbox of Assessment Methods© that is now a guide used in assessment of physicians in training (Bashook & Swing, 2000) and other sources (Bashook, 1994). Additionally, the methods can be grouped into four assessment categories according to what each is best at measuring: (1) assessment of knowledge, (2) assessment of decision-making, (3) assessment of practice performance and personal attributes, and (4) assessment of skills and tasks.

Assessment of Knowledge

This usually refers to assessing recall of facts, concepts, principles, and basic application in a standard examination format. There are three common exam formats: multiple choice questions (MCQs), essay questions, and short-answer questions.

TABLE 1
Assessment Methods and Recommended Uses^a

<i>Assessment Focus</i>	<i>Method</i>	<i>Description</i>	<i>Recommended Use</i>
Knowledge	Written exam—multiple choice questions	Many test questions in multiple choice with single response or matched answers	Assess knowledge of facts and concepts
Knowledge	Written exam—essay questions	Present challenging problem or question ask for written essay response	Assess intellectual synthesis of ideas, comprehension
Knowledge	Written exam—short answer questions	Present challenging problem or question and response in few words	Assess knowledge of facts and concepts, comprehension
Decision-making	Oral exam (standardized)	Case-based written, simulated, live patients, own case reports as focus of examiners' questions	Assess case specific decisions, know own limits
Decision-making	Key features cases	Written case focus on essential client management decisions	Assess decision-making different stages in client care
Performance and attitudes, skills/tasks	Global ratings	Raters judge general abilities usually retrospective after repeated observations	Assess performance on list of competencies synthesized by rater or multiple raters
Performance and attitudes	Supervisor's narrative reports	Documented summary of supervisor's judgments about person	Assess synthesized judgments about abilities, limitations

Performance and attitudes	Client surveys	Survey questions to clients asking about care satisfaction, judgments about providers, facilities, other services	Accumulated forms assess patterns or incidents, attitudes, actions, communications, professionalism
Performance and attitudes	Client record review	Use criteria and protocol to judge documented care	Assess decision-making, care follow-through
Performance and attitudes	Portfolios	Predefined expected products of practice, reflect about quality and impact	Assess judgments about quality, master professional and intellectual behaviors
Performance and attitudes	360-degree evaluation	Survey forms completed by multiple people in person's sphere of practice	Assess perceived judgment, professional behavior, communication skills
Performance and attitudes and skills/tasks	Standardized patient simulations (SPs)	Actors trained to follow protocol and simulate a client with clinical condition	Assess interview, communication, therapeutic alliance, interpersonal skills
Skills/tasks	Simulations and models	Imitate clinical situations by computerized and live cases	Assess attitudes, decisions, skills, abilities, and training
Skills/tasks	Checklist/rating scales	Ratings on specific behaviors, activities, tasks, or sequence of actions	Assess specific observable behaviors or actions

^aModified from ACGME Toolbox of Assessment Methods© (Bashook & Swing, 2000).

Multiple Choice Questions (MCQ). The typical standardized test that contains hundreds of questions often presents a brief synopsis of a client situation. The candidate is to select the best answer among four or five options. The individual taking the exam is judged by how many of the preferred responses are chosen. Questions are scored as correct or incorrect and tallied to decide a pass/fail decision or rank the person among peers. The questions are selected from a pool of questions based on a test blueprint that defines the content to be assessed. Experts on the content pre-select the correct answers. When properly designed, this type of written exam is considered the gold standard in knowledge assessment. Nearly all members of the behavioral health workforce are expected to pass standardized written examinations in the multiple-choice format at some point in their career.

These written exams are typically developed and administered by a certifying or licensure board. The MCQ exams are administered on paper or delivered on a computer as one or more half-day sessions, with around 150–200 questions per session. Some boards have one or even two full days of exams (300–600 test questions per exam). Well-constructed exams comply with accepted psychometric standards for reliability and validity (reliability can be as high as 0.98 for a diverse group of candidates). Credibility of results is high by all who rely upon test scores as evidence of the candidate's knowledge. Although expensive to create and administer, it is quite feasible to use this format for large-scale national testing of candidates.

Training instructors often assumes that constructing quality written exam questions will be easy. The research evidence suggests these instructor-made tests are rarely reliable (e.g., too few questions), and may not cover the content adequately (e.g., questionable validity). Also, design flaws with the MCQ technology contribute to unreliable scores. For example, one question gives hints to help the less capable answer other questions, or the questions contain grammatical errors that guide more astute test-takers (Case & Swanson, 2003; Joint Committee on Testing Practices, 1999).

Essay Questions. Essay questions present the test-taker with a challenging problem or scenario and ask him/her to explain how s/he would address the problem or scenario in a written essay response. Lengths of allowable responses can vary, and scoring is completed by content experts. The grading may be pass/fail or use various rating scales. Issues of reliability surface when multiple graders judge performance or one person must grade many essays. Reliability can be improved by training and monitoring the graders. The Educational Testing Service has

developed software to automate grading of essays and short-answer questions (Educational Testing Service, 2004).

Continuous quality improvement is a newer technology that some suggest could be used to measure practice performance.

Short Answer Questions. When using a short-answer question format, a brief synopsis of a client situation or problem is presented and the person responds with a phrase or one-sentence answer. Experts on the topic score answers. Grading answers can be automated using computer software (Educational Testing Service, 2004), which limits problems of inter-judge reliability. Short-answer questions are often used in written exams for limited numbers of trainees in place of the MCQ format because they are much easier to construct and do not require sophisticated technology to score.

Assessment of Decision-Making

At every stage in care, the practitioner must make judgments about critical actions that can affect a client. Decision-making and judgment cannot be assessed with standardized MCQs. They require assessing the use of knowledge in realistic practice situations (Page, 1995). The following assessment methods are effective for assessing decision-making if designed and used appropriately: case-based oral exams and key features cases.

Case-Based Oral Exams. This technology is used extensively in certification examinations for psychiatry (Juul & Scheiber, 1994), psychology, including specialties in psychology (see American Board of Professional Psychology, <http://www.abpp.org>), and other behavioral health disciplines requiring a professional degree. The candidate can be presented with case material in a variety of formats: written vignettes, images, their own client case reports, or live client situations. As the case unfolds, the candidates must explain their decisions about assessment, diagnoses, treatment planning, and/or managing the case. Examiners can question candidates on reasons for their decisions. Adding hypothetical variations to the presenting case tests the candidate's limits of expertise and actions they would take once those limits are reached (Mancall & Bashook, 1995). A typical examination lasts 30–60 min per session, with four to eight sessions. In this time frame, a well-constructed exam can question a candidate on 12–36 cases, and obtain from 50 to 100 measures of clinical decision-making. Estimated score reproducibility (reliability) has been

consistently above 0.90 for well designed and administered oral exams for certification (Lunz, 1995; Lunz et al., 1994).

Quality oral exams require extensive training of examiners (Des Marchais & Jean, 1993; McDermott et al., 1991), standardization of cases, pre-established scoring schema, and careful monitoring of administration to obtain reliable and valid results (Bashook, 2003; Mancall & Bashook, 1995). When designed properly, the case-based oral examination is a good predictor of practice performance (Solomon, Reinhart, Bridgham, Munger, & Starnaman, 1990).

Key Features Cases. This approach is a written examination where the person must make decisions for critical actions (key features) occurring at various stages in the case. Experts score responses based upon previously established criteria. Each case is counted as a single score (Page, Bordage, & Allen, 1995). Key features cases are currently used in physician licensure examinations in Canada (Page et al., 1995). This method has not been used in assessments of clinicians in the behavioral health workforce, but certainly could be incorporated into written exams during training and practice.

Assessment of Practice Performance and Personal Attributes

Assessing the performance of trainees involves assessments of observed behavior with clients over time, or in specific observed client encounters. Most commonly used methods are: global rating forms, supervisor's summary reports, client surveys, client record audits, portfolios, and 360-degree evaluations.

Global Rating Forms. A rater uses a form with multiple categories of performance to provide retrospective impressions/judgments about a person's performance. The rater can not only incorporate observed performance over time, but often include a synthesis of second-hand information from multiple sources. The rating scales usually include a place for judgments about overall competence and space for written comments. Scoring global rating forms includes separate tallies of rating scales with averages, frequency counts, the ratings by multiple raters, and qualitative evaluation of comments. There is some evidence that global ratings are superior for assessing performance compared to checklists (Regehr, MacRae, Reznick, & Szalay, 1998). This assessment method is used frequently in supervised clinical care situations, with supervisors or more senior practitioners rating junior practitioners or trainees. It is used in all behavioral health training programs leading to professional degrees and for behavioral health specialists.

Supervisor's Summary Reports. These reports are summaries produced biannually or annually (for employment), and provide documentation of a supervisor's evaluation of trainees or practitioners employed in a behavioral health facility. They serve as a compilation of the supervisor's judgments about the competencies of the person accumulated over months or a year. Often the reports are confidential. This report is ubiquitous, and used in both training and practice for all levels of the behavioral health workforce.

Client Surveys. Clients complete a questionnaire about specific encounters with a practitioner, the setting, and related care issues. Typical assessments include satisfaction with care, overall quality of care, competencies in interpersonal relations, therapeutic relationships, perceived expert knowledge, and professional practices. Accumulated across a number of clients, the summary of results and highlighted incidents (positive and negative reports from clients) can provide insight into how clients perceive a practitioner's professional demeanor, attitudes, and care (Weaver, Ow, Walker, & Degenhardt, 1993). Scoring is done by experts comparing findings against expected performance at the level of training and circumstances of practice situation.

Client Record Audits. This approach is customarily used to assess performance in practice with trained auditors performing a confidential review of case records and judging findings based on previously defined protocols and criteria. Audit information from multiple cases is easily converted into statistical descriptions to measure compliance with expected practices. Scores are useful for identifying strengths and weaknesses in practice performance when compared to similar practitioners. Some medical specialty certifying boards have introduced client record audits as part of the re-certification for their specialty (Bashook, 1994).

Roe points out, "A high level of competence is a prerequisite for good performance; it does not guarantee adequate performance."

Portfolios. The portfolio is a defined collection of products prepared by the student or practitioner that demonstrates progress in learning about or mastery of a competency. Products can be from training or practice experiences (e.g., clients encountered, ethical situations). For each product required in the portfolio, there are specifications based on what competencies will be assessed. In addition, the trainee or practitioner might be required to prepare a statement reflecting upon quality of the product, what was learned, and assessment of current competency.

Portfolios have been used to assess psychiatrists during residency training on attitudes, professionalism, and experience-related competencies that are not easily and systematically measured by other means (O'Sullivan, Cogbill, McClain, Reckase, & Clardy, 2002). Supervisors and instructors can score the portfolio against pre-determined standards. When properly designed, portfolios can be a reliable method to assess the more intangible attributes of competence, even in high-stakes assessments (Roberts, Newble, & O'Rourke, 2002).

360-Degree Evaluations. Often used in business, 360-degree evaluations are multiple ratings done retrospectively, concurrently, and separately by people in the sphere of influence of the person being evaluated (e.g., supervisors, colleagues, subordinates, clients, referring clinicians). All raters receive the same written survey containing rating scales and requesting judgments about a person's performance for a specific time period. The raters are strongly encouraged to add comments that illustrate the reasons for the ratings. Competencies often assessed include the person's clinical performance, interpersonal relationships, teamwork, knowledge application, communication skills, attitudes, and professionalism (Hall et al., 1999). The rating scales can be tabulated to produce a numeric score, and comments are organized to provide insight into the raters' perceptions about the person. A variation on the 360-degree evaluation is multiple peer ratings of performance that emphasize only the attributes that each peer is best at rating (Ramsey et al., 1993).

Using the 360-degree report requires caution in keeping information confidential, because comments are often sensitive, and exposure can be detrimental. This assessment method is used most effectively during employment situations for individuals who have some supervisory responsibilities, or training situations where the person has a significant role in team care.

Assessment of Skills and Tasks

Competencies involving specific skills or actions in client assessment, treatment, or care management can be assessed individually both in the context of care and during training. In order to give the skills or tasks a context, the assessment requires the presentation of a clinical case situation, even if only a brief description of the patient's characteristics. More elaborate case situations are used when the assessment attempts to mimic the realities of clinical practice as much as possible, and these commonly use role-playing simulations with live interactions or computers to create virtual reality environments. Typical assessment methods are: rating scales and checklists, role-playing computer simulations, and role-playing standardized patient examinations.

Rating Scales and Checklists. Rating scales and checklists are used during live or videotaped observations of a trainee or practitioner as a means of guiding the evaluation, and as documentation of what was observed. These assessment methods are very similar in how they are used, but differ in one respect. For checklists, the rater decides if the person being evaluated has or has not performed a specific action. If performed, the rater then checks the appropriate box on the form. With rating scales, the rater may judge not only completing a task, but also how well it was performed along a spectrum of excellent to poor or other range of quality performances. The additional step of judging the quality of a performance introduces greater variability in the ratings due to differences in interpreting the meaning of scale descriptions (e.g., what exactly does excellent or average mean). Personal biases about what behaviors should count more or less also influence the consistency of ratings across raters are, along with a tendency of raters to differ about how severe or easy they are when grading another's performance (rater severity). These variations in judgment are one reason rating scales may have a lower reliability than checklists, unless the rater is trained how to use the scales. There are statistical methods to correct for rater severity (Lunz et al., 1994). Also, training of raters improves consistency and validity of the raters' judgments (Winckel, Reznick, Cohen, & Taylor, 1994). It appears that global rating scales may provide more reliable measures of performance compared to checklists when the tasks are complex (Regehr et al., 1998). Typical uses include: completing a series of steps in a client workup such as mini-mental health status, or assessing completion of steps in a protocol for planning a client's discharge from a restricted care unit.

Role-playing Computer Simulations. Simulations used in assessment closely resemble reality. The focus is on the essential realistic clinical problems to be solved, while stripping away irrelevant distractions (Clyman, Melnick, & Clauser, 1995). Successful simulations, whether on computer, paper-and-pencil, or through role-playing, force the trainee or practitioner to sort through a wide variety of options as they clarify the important clinical problems and challenges to address and attempt to solve the problems. Simulations on computer have been developed to train surgeons, anesthesiologists, and other procedure-oriented doctors to manage new invasive technology like arthroscopy (Taffinder, Sutton, Fishwick, McManus, & Darzi, 1998).

Life-sized computerized adult and child mannequins have been used in an operating room simulation to train anesthesiologists in basic and advanced anesthesia treatments, including crisis situations (Gaba et al., 1998). These technologically advanced simulations, referred to as virtual reality (VR) environments, are commercially available at a cost of around

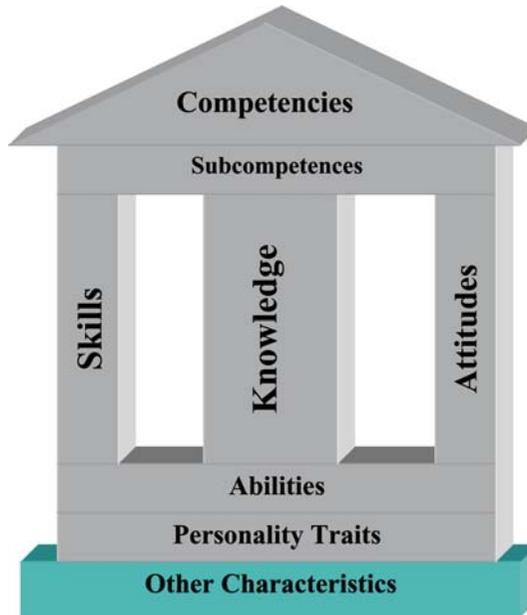
\$50,000 each, and include case-based software. There are additional expenses for creating tailor-made case scenarios, maintaining equipment and space, and employing technical staff to run the simulations. Some medical schools and hospitals have purchased VR equipment to train clinical staff, medical students, residents, and physicians. It is anticipated this technology will be widely adopted in medical curricula, because there are fewer opportunities to learn directly from patients and clients.

Role-playing Standardized Patient Examinations. During a standardized patient examination (SP), a trainee or practitioner is presented with a realistic scenario and must interact with a live person (the SP) in a role-playing simulation, as if the SP were a real client. The SP is an actor who has been previously trained to simulate a client with a realistic condition and appropriate emotional state. The trainee or practitioner's performance during the encounter can be evaluated against expected competencies defined in advance, and documented either by the SP or an observer. The SP encounter can last 10–30 min, followed by at least 10 min for the SP or an observer to rate the performance. Frequently, the encounters are observed and videotaped to protect the SP and the person being evaluated. SPs are widely used in training medical students and physicians in training, and for continuing medical education experiences (Guagnano, Merlitti, Manigrasso, Pace-Palitti, & Sensi, 2002). The SP examinations are most effective to evaluate the following competencies: workup/assessment of a client (medical, social, emotional, or other history, physical examination skills); communication skills, including giving bad news and counseling patients; and managing complex situations that could harm patients or staff when mishandled (e.g., suicidal patient, aggressive client behavior, paranoia).

Conceptual Frameworks for Assessment in Training and Practice

A distinction needs to be made between assessing competence and assessing performance in practice. Competence to practice is measured during training in controlled situations, at the time credentials or licenses are obtained, through objective examinations (written and oral). Assessment of performance during actual practice is measured either with assessments that are a snapshot of client care, or accumulated assessments over time (somewhat like video clips of practice with annotated ratings of the performance quality). Statistical analyses discern useful trends and outlying behaviors for improving quality of care, and look for patterns in the setting that need quality assurance interventions. In either the snapshot or video format, they are direct measures of practice, not implied capacities based on exams. Roe (2002) described a tradi-

FIGURE 1
Competence Architecture Model (Roe, 2002)



tional approach for assessment of psychologists' competencies prior to practice (during training) that is applicable to any occupation. Roe's model, the "competence architecture model" (2002), was intended as a guide for incorporating defined competencies for curricular design and program accreditation, but it works equally well for assessing competencies of anyone in the behavioral health workforce.

The model proposed by Roe can be visualized as a Greek temple (see Figure 1). He depicts expected competencies capping a building that has foundation layers of abilities, personality traits, and other personal attributes, all potentially measurable by assessment methods. Pillars of acquired learning are the traditional KSAs (knowledge, skills, and attitudes) where depth and breadth of learning are assessed. Practical learning supports the roof during supervised training. The knowing how and when that integrates the KSAs with the foundation layers become subcompetencies. Subspecialties combine KSAs with other abilities and personal attributes, all of which work together when performing a specific and demonstrable part of the clinical care. Typical subcompetencies include the evaluation of a client or the articulation of a treatment plan for a client. The roof of the model is made up of the competencies essential to practice. By combining assessments for the architectural ele-

ments below the roof of competence, one can infer whether a person has the appropriate competencies essential to practice.

The individual competencies defined by the Annapolis Coalition (Hoge et al., in press) are equivalent to subcompetencies in Roe's competence architecture model. In assessment, the preference is to measure each subcompetency separately and accumulate the results to make judgments about a person's overall competence. For example, a subcompetency is the ability to perform an appropriate and focused intake interview with a client and/or family.

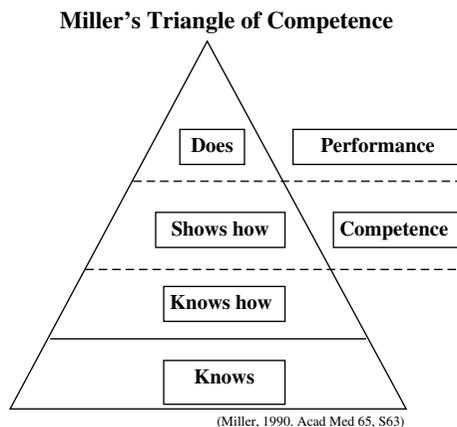
A variety of assessment methods could be used to generate an aggregate score composed of multiple measurements of this ability. A systematic assessment could include a knowledge test about essential steps and theory in history taking, live or simulated observations of the student or practitioner interviewing a client, documented accumulated ratings on intake interviewing skills observed and judged by faculty, supervisors or senior trainees over weeks or months of supervised practice, and measures of the person's attitudes about clients' cultural differences assessed using validated attitude scales. An aggregate score that combines these measures would require adjustments for statistical reliability of each measure. Interpreting the score must be tempered by qualitative adjustments for the person's communication style, personality attributes, assumed relative validity of each measure, and limits and circumstances when each measure was taken.

Accumulating valid measures for each subcompetency is essential, but as Roe points out, "A high level of competence is a prerequisite for good performance; it does not *guarantee* adequate performance." This model provides the framework used in this paper when explaining how to design assessment of competencies for entry into practice.

Miller's Triangle (1990) provides a useful framework for structuring assessment of performance in practice. The triangle is like an inverted pyramid, with four progressive stages of assessment: "knows," "knows how," "shows how," and "does" (see Figure 2). All four stages clearly define progressive capabilities, and build on abilities in the lower stages. Also, Miller's Triangle visualizes the well-established principle that assessment of a person's knowledge is important, but not sufficient to predict they will apply the knowledge in practice (Kennedy, Regehr, Rosenfield, Roberts, & Lingard, 2003).

Considering the roles, responsibilities, and settings of behavioral health practice requires adding two more components to the Miller Triangle: (1) systems-related influences on practice (e.g., facility-specific regulations, policies, patient expectations, governmental regulations, and access to other health professionals), and (2) individual-related influences (e.g., mental and physical health of the practitioner, relationships

FIGURE 2
Miller's Triangle of Competence Assessment (Miller, 1990)



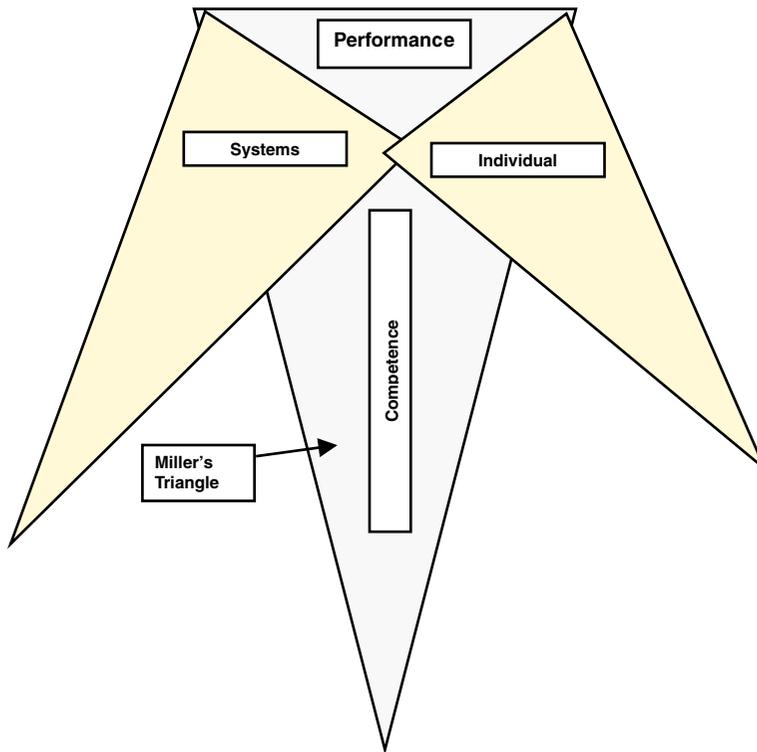
with others like patients, other practitioners, and their family, and state of mind at time of performance practice assessment). Rethans and colleagues (2002) refer to this more complex model of assessment and competence as the “Cambridge Model,” after the conference where it was proposed (see Figure 3).

Measurement of practice performance is complex because of variability in forces external to the individual. The Institute of Medicine (2000, 2001) reported about safe health care, and emphasized that quality performance by individual practitioners depends directly upon the health care systems where they work. The following is a short list of the common systems-related factors that can influence practice performance, and must be considered when interpreting results from an assessment program for behavioral health providers:

- Case mix and quantity of clients
- Differences in priority setting by individuals
- Institutional policies and regulations
- Legal, ethical, and other limits in how one can practice
- Expertise and teamwork of available clinical team members
- Options for referral to practitioners with greater expertise
- How care will be paid for and limitations in insurance coverage

Also, the validity of predicting quality of practice performance from objective assessments like exams for re-licensure or renewal of certification (the “knowing how” in the Miller model) depends to a large extent

FIGURE 3
The Cambridge Model for Assessment of Performance
 (Rethans et al., 2002)



upon how well these assessment methods are adjusted to account for the variabilities inherent in daily practice.

Looked at from the perspective of the practitioner, maintenance of competence in practice requires a purposeful self-directed learning agenda that combines opportunities to participate in continuing education activities and on-the-job learning. Often, client needs and expectations, as well as a wish to manage care situations most effectively, drive most practitioners' learning agendas (Bashook, 1993). A method to assess this competence in practice has not been well developed.

Another product of the Cambridge Conference meeting was a framework for implementing a practice performance assessment program (Lew et al., 2002). The proposed framework outlines three broad domains to address in planning, and specifies the questions and decisions to consider. It offers guidance for creating defensible procedures that should address the concerns of all stakeholders. The domains are:

1. *Purposes and Outcomes.* What are the purposes of the assessment? Whose purposes are being met? Are they unambiguously stated and made available prior to implementing the program?
2. *Planning the Practice Assessment Program.* What steps are taken to assure fairness and defensibility of the process and results? Is the plan clearly described, including who are the assessors, what methods are used, what is known about the technical characteristics of the methods?
3. *Processes.* How will the program be administered and communicated to the stakeholders and assessors? How will methods be developed and used? What are the security issues? What are the policies and rules regarding the amount of time for the assessments and appeals procedures? What are the feasibility issues like cost and resource needs to produce credible and useful performance data?

The framework is partly based upon the extensive experience of the United Kingdom's General Medical Council peer review program, which assessed physicians' practice performance. The UK program withstood court litigation to remove doctors' licenses (Southgate et al., 2001), and uses a portfolio method combining interviews, tests of competence, and self-reports, which generates more than 700 judgments about the doctor's practice performance.

A less ambitious suggestion for assessing performance practice is to provide tools for the individual practitioner to create a self-directed learning portfolio (Roberts et al., 2002) and report progress to certifying and licensure boards using web-based software (Bashook & Parboosingh, 1998). This approach would fit into one component of the maintenance of certification programs by the physician specialty boards in the U.S. and Canada (American Board of Medical Specialties, 2003; Royal College of Physicians and Surgeons of Canada, 2004). It also takes into consideration on-the-job learning directly related to personal and institutional influences the practitioner brings to the workplace (see Figure 2). It is important to realize that some practice roles cannot be assessed with any assurance until the person has begun working in a practice setting and has opportunities to demonstrate their capabilities over time, under real conditions (Cunnington & Southgate, 2002).

Accumulation of continuing education credits in place of direct performance assessments has no value when assessing practitioners' maintenance of competence.

Continuous quality improvement is a newer technology that some suggest could be used to measure practice performance (see Institute for Healthcare Improvement, <http://www.ihi.org>). It is based on the principles of quality control used in engineering systems as adapted to human behavior and health systems. Most recently, quality improvement initiatives have focused on patient safety themes, which supports the Institute of Medicine report about “errors in medicine” and health care system deficiencies (Institute of Medicine, 2000, 2001). The assessment pays direct attention to individual behaviors that are influenced by the systems where they work, which in turn influence quality. It seems to work in medical settings with defined expectations for patient care decision-making and outcomes.

Once in practice, the person may have the competence and know-how and perform admirably when the opportunities arise, yet still have few situations to perform all they can do. Demonstrating pre-practice competence does not necessarily mean the person will find him or herself in a practice environment designed to support competence, and so may not function competently in practice. The reality of practice places constraints on how competencies are routinely used, and the practice setting adds additional restrictions that necessitate conformity to team preferences or institutional policies and practices, whether or not these preferences, policies, or practices have as a basis empirical knowledge.

These variations in settings, roles, and responsibilities will influence the individual practitioner’s abilities to maintain the initial competencies assessed at entry into practice. Complicating the equation are the growing trends that require practitioners to demonstrate continuing or maintenance of competence by periodic reassessments for re-registration of a license or renewal of certification (Bashook & Parboosingh, 1998). These reassessments often occur at intervals of two or three years for licensure, and 5–10 years for renewal of certification. It’s important to recognize that accumulation of continuing education credits in place of direct performance assessments has no value when assessing practitioners’ maintenance of competence (Cunnington & Southgate, 2002). An alternative is to adopt the maintenance of certification programs being implemented in Canada and the United States (American Board of Medical Specialties, 2003; Royal College of Physicians and Surgeons of Canada, 2004).

Some behavioral health providers without professional degrees, or advanced certified training do not have these re-registration and renewal requirements. However, all groups are reassessed for continuing competence through employment evaluations, practice opportunities, and attempts to advance through adding specialized expertise with additional certifications.

Recommended Best Practices in Assessment of Providers

In considering which methods to adopt, it is important to realize that no single assessment method can evaluate all competencies, and more than one method may measure the same competencies (see Table 1). Ideally, the best approach is to develop an assessment blueprint that identifies multiple assessment methods tailored to the competencies to be measured, and accounts for feasibility of using the methods when considering the career stage of a practitioner (year in training, or practice roles in clinical settings). An example is a peer assessment program for family physicians in Canada that uses written exams, case-based and chart-stimulated oral exams, and standardized patient cases (Norman et al., 1993).

Schuwirth and colleagues (2002) proposed guiding principles that would combine practice performance assessment methods into results that all stakeholders would consider coherent, credible, and defensible. In their view, the combination of assessment methods should provide a whole portrait of the practitioner. The essential ingredients include: having large samples of behavior to assess, irrespective of the assessment methods used; organizing the sequence and intensity of assessments into a structure, but not an overly regimented or prescriptive structure; and using multiple assessment methods to reduce risk of bias due to any one method. Also, it's important to keep in mind that ways of assessing competencies are not static, and need to be revised to be consistent with current priorities in the discipline, public expectations, current scientific knowledge, and improvements in assessment methodology.

With these caveats noted, the following are some suggested best practices citations that build upon the published literature (see American Board of Professional Psychology, <http://www.abpp.org>; Bashook, 1994). For examples of assessment practices with other health care providers, see Landon, Normand, Blumenthal, and Daley (2003); Browning et al. (1996); Swanson et al. (1995); and Foulkes et al. (1993). These recommended best practices are grounded in the conceptual framework for assessments in training, the "competence architecture model" (Roe, 2002); and the framework for assessment in practice, the "Cambridge Model of Assessment" (Rethans et al., 2002). All suggestions are tempered by considerations of the reliability, validity, feasibility, and credibility of the assessment methods.

Within each of the traditional behavioral health disciplines, there are templates for assessment practices, some much more detailed than others. This is also true for some practice areas that have traditionally put less emphasis on academic credentials and more on life experiences, such as addictions counseling and the newly created peer support spe-

cialist category. There are also educational programs being developed targeted towards families and primary consumers, for which assessment strategies are in their earliest stages of development. Readers seeking detailed information should access professional association websites or seek information related to intervention strategies with specific population targets (e.g., assertive community treatment for persons with serious and persistent mental illnesses).

Best Assessment Practices: Professional Degreed Practitioners

The Example of Psychiatrists. The medical student who plans to enter psychiatry after completing the M.D. degree is continuously evaluated over the four-year medical school curriculum in a carefully constructed and progressive assessment process that resembles the competence architecture model. All accredited medical schools in the U.S. and Canada must have defined graduation competencies and a comprehensive system of evaluating medical students (see <http://www.lcme.org/standard.htm>). After graduating medical school, assessments for residents in psychiatry for four years (general psychiatry) or five to six years (child and adolescent psychiatry) shift emphasis from evaluating knowledge and basic clinical skills and tasks to evaluation of core psychiatric competencies (Scheiber, Kramer, & Adamowski, 2003). The accumulated results of these assessments during residency determine whether the graduate is qualified to become a candidate for certification by the American Board of Psychiatry and Neurology. Advanced certification after training in child and adolescent psychiatry and other psychiatric specialties involves a similar two-stage assessment process.

Best Assessment Practices: Trained Therapists with College Degrees

The Example of Creative Arts Specialists. The Art Therapy Credentials Board (2004) has developed a certifying process for art therapists that includes training requirements and a written case-based knowledge examination. The exam uses MCQ items with cases to cover the six major content domains of the discipline: (1) psychological and psychotherapeutic theories and practice, (2) art therapy assessment, (3) art therapy theory and practice, (4) recipient populations, (5) art therapy media, and (6) professionalism and ethics. Assessments of performance in practice would greatly enhance the credibility of the certificates. These assessments could be obtained at reasonable cost and effort through systematic reports using a portfolio assessment method.

Best Assessment Practices: Non-Degreed Staff

Example of Certification in Alcohol and Other Drug Abuse. The International Certification and Reciprocity Consortium/Alcohol and Other Drug Abuse (2004) has established certification standards for alcohol and other drug abuse counselors. Most American states and more than a dozen countries have adopted these standards. The training requirements include 240 h of formal didactic instruction in workshops, courses, institutes, in-service, and distance education. Supervised practical training requires 300 h, covering 12 core functions with assessment of targeted skill development and demonstrated application of knowledge within an alcohol or drug counseling setting.

In addition, entry to certification requires 6000 h (three years) of supervised experience in providing alcohol or other drug abuse counseling services. An associate's degree and other behavioral science course work can substitute for some of these training and course requirements. Besides the reports of successfully completing the supervisor's evaluations, the candidate must pass a written examination (MCQ format) designed by a team of international experts in alcohol and substance use. Finally, a case-based oral examination (write-up of a single client that the candidate has managed) must be passed. Peers who have advanced certification evaluate this case in a structured oral examination.

Clearly, this certification program closely follows the "competence architecture model," having the counselor build competency with didactic foundational course work, plus extensive focused and supervised skill and practice development, in addition to supervised training in practice.

CONCLUSION

The recommended best practices for assessment of the behavioral health workforce can be summarized in the following guidelines:

1. Define the content and competencies to be assessed in an assessment plan or blueprint as the first step in creating a valid assessment program.
2. Provide evidence that the implemented assessment methods measure what was intended in the plan with supporting data, statistical analysis, and logical explanations. The assessment evidence should:
 - Assure that the assessment is reliable, showing the amount of error or variability that could occur if the same assessment were repeated with the same group of trainees or practitioners.

- Present accumulated evidence of the validity of assessment results for a specific group of people in specific circumstances to demonstrate that the results can be interpreted to measure what they are purported to measure. It is the scores on the assessment, not the method, that is valid.
 - Demonstrate the feasibility of an assessment method with realistic estimates of cost in time and effort to develop, test, implement, and obtain valid results when using the method.
 - Demonstrate the credibility of an assessment where all stakeholders who rely upon the assessment results consider the methods and the findings plausible, consistent, and useful for the intended purposes.
3. Use the “competence architecture model” (Roe, 2002) as a guide for combining assessment methods appropriate for evaluating trainees during training or at the completion of training (e.g., initial certification or licensure).
 - Assessments used during training for purposes of feedback for trainees do not need the same high reliability and rigorous validity standards as in high-stakes assessments such as those involving credentialing and licensure.
 - Assessments for licensure and certification (initial and renewal) should include a well-engineered blueprint and evidence of validity and reliability that is credible to defend against challenges.
 4. Use the “Cambridge Model” (Rethans et al., 2002) as a guide for combining assessment methods appropriate for evaluating performance in practice (e.g., continuing quality improvement of practice, renewal/maintenance of certification, re-registration of license).
 - Sample multiple behaviors and practice events using a variety of assessment methods.
 - Avoid overly structured assessment program that trivialize what is to be assessed (Schuwirth et al., 2002).
 5. Construct new assessment methods using the following sequence: (1) content and testing experts work together to develop the new method to assure content accuracy and technical integrity, (2) pilot test and revise assessment cases or test items as needed, and (3) perform psychometric analyses of results every time the methods are used (Browning et al., 1996).

There is a growing global emphasis on assessing the competence of all health care providers, especially physicians, as reflected in standards for accreditation requiring assessment of competencies (World Federation

for Medical Education, 1998). This trend continues into initial specialty certification, with time-limited certification that requires physicians to renew their certification through a process called “maintenance of certification/competence” (Bashook & Parboosingh, 1998; Cunnington & Southgate, 2002). This practice is the norm in medicine throughout North America (American Board of Medical Specialties, 2003; Royal College of Physicians and Surgeons of Canada, 2004), and is rapidly taking hold in Europe and other regions of the world (see European Union of Medical Specialists, <http://www.uems.net>). It is common for trends that start in medicine to influence other health care disciplines. Therefore, assessment plans, which demonstrate maintenance of competence, are soon likely to be an important priority for all behavioral health disciplines. The medical model of competence and performance assessment is one option, but the behavioral health workforce should consider alternatives tailored to their specialized roles, responsibilities, and settings.

A start could be periodic cross-disciplinary meetings to exchange information and experience about assessment programs. Also valuable would be a grant funding mechanism to foster creating better assessment tools and methodology specific to common competencies in behavioral health care. No matter how this effort is achieved, building and using quality assessment methods will not occur without significant planning, support, and cooperation among all who have a stake in behavioral health care.

REFERENCES

- American Board of Medical Specialties. (2003). *ABMS annual report and reference handbook*. Available from: <http://www.abms.org>.
- Andrich, D. (1988). *Rasch models for measurement*. Newbury Park CA: Sage Publications.
- Art Therapy Credentials Board. (2004). *Art therapy credentials board: Booklet of information and study guide for certification examination*. Available from: <http://www.atcb.org/>.
- Bashook, P.G. (1993). Clinical competence and continuing medical education: Lifelong learning to maintain competence. In C. Coles & H.A. Holm (Eds.), *Learning in medicine* (pp. 21–41). Oslo, Norway: Scandinavian University Press.
- Bashook, P.G. (1994). Beyond the traditional written and oral examinations: New certification methods. In J. Shore & S. Scheiber (Eds.), *Certification, recertification, and lifetime learning in psychiatry* (pp. 117–138). Washington, DC: American Psychiatric Press.
- Bashook, P.G. (2003). *Structured case-based orals are a valid and reliable measure of physician's clinical decision-making*. Chicago, IL: American Educational Research Association.
- Bashook, P.G., & Parboosingh, J. (1998). Continuing medical education: Recertification and the maintenance of competence. *British Medical Journal*, *316*, 545–548.
- Bashook, P.G., & Swing, S. (2000). *Toolbox of assessment methods*©, version 1.1. Evanston, IL: Accreditation Council for Graduate Medical Education /American Board of Medical Specialties. Available at <http://www.acgme.org>.
- Brennan, R.L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295–317.
- Browning A.H., Bugbee A.C., & Mullins M.A. (Eds.) (1996). *Certification: A NOCA handbook*. Washington, DC: National Organization for Competency Assurance.
- Case, S.M., & Swanson, D.B. (2003). *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners. Available at <http://www.nbme.org>.

- Clyman, S.G., Melnick, D.E., & Clauser, B.E. (1995). Computer-based case simulations. In E.L. Mancall & P.G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139–149). Evanston, IL: American Board of Medical Specialties.
- Cunnington, J., & Southgate, L. (2002). Relicensure, recertification, and practice-based assessment. In G.R. Norman, D.I. Newble & C.P.M. Vander Vleuten (Eds.), *International handbook of research in medical education* (pp. 883–912). Amsterdam: Kluwer Academic Publishers.
- Des Marchais, J.E., & Jean, P. (1993). Effects of examiner training on open-ended, high taxonomic level questioning in oral certification examinations. *Teaching & Learning in Medicine*, 5(1), 24–28.
- Educational Testing Service. (2004). *Graduate record examination (GRE)*. Available from: <http://www.gre.org/tiindex.html>.
- Foulkes, J., Bandaranayake, R., Hayes, R., Phillips, G., Rothman, A., & Southgate, L., (1993). Combining components of assessment. In D. Newble, B. Jolly & R. Wakeford (Eds.), *The certification and recertification of doctors: Issues in the assessment of clinical competence* (pp. 134–150). Great Britain: Cambridge University Press.
- Gaba, D.M., Howard, S.K., Flanagan, B., Smith, B.E., Fish, K.J., & Botney, R. (1998). Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology*, 89, 8–18.
- Georgia Certified Peer Specialist Project. (2004). *Georgia Certified Peer Specialist Project*. Georgia Mental Health Consumer Network, Division of Mental Health, Developmental Disabilities and Addictive Diseases. Available from: www.gacps.org.
- Guagnano, M.T., Merlitti, D., Manigrasso, M.R., Pace-Palitti, V., & Sensi, S. (2002). New medical licensing examination using computer-based case simulations and standardized patients. *Academic Medicine*, 77(1), 87–90.
- Hall, W., Violata, C., Lewkonja, R., Lockyer, J., Fidler, H., & Toews, J., et al. (1999). Assessment of physician performance in Alberta: The physician achievement review. *Canadian Medical Association Journal*, 161(1), 52–57.
- Hoge, M.A. & Morris, J.A. (Eds.) (2002). Behavioral health workforce education and training. [Special issue]. *Administration and Policy in Mental Health*, 29(4/5), 297–303.
- Hoge, M.A. & Morris, J.A. (Eds.) (2004). Implementing best practices in behavioral health workforce education: Building a change agenda [Special issue]. *Administration and Policy in Mental Health*, 32(2), 83–205.
- Hoge, M.A., Tondora, J., & Marelli, A.F. The fundamentals of workforce competency: Implications for behavioral health. *Administration and Policy in Mental Health*(in press).
- Institute of Medicine. (2000). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.
- Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st Century*. Washington, DC: National Academy Press.
- International Certification & Reciprocity Consortium/Alcohol & Other Drug Abuse. (2004). *Standards for certified alcohol and other drug abuse counselors*. Available from: <http://www.icrcaoda.org/>.
- Joint Committee on Testing Practices (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association National Council on Measurement in Education.
- Juul, D., & Scheiber, S.C. (1994). The part II psychiatry examination: Facts about the oral examination. In J.H. Shore & S.C. Scheiber (Eds.), *Certification, recertification, and lifetime learning in psychiatry* (pp. 71–90). Washington, DC: American Psychiatric Press.
- Kennedy, T., Regehr, G., Rosenfield, J., Roberts, W., & Lingard, L. (2003). *Degrees of gap between knowledge and behavior: A qualitative study of clinician action following an educational intervention*. Chicago, IL: American Educational Research Association.
- LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions*, 17(2), 178–197.
- Landon, B.E., Normand, S.L., Blumenthal, D., & Daley, J. (2003). Physician clinical performance assessment: Prospects and barriers. *Journal of the American Medical Association*, 290(9), 1183–1189.
- Lew, S.R., Page, G.G., Schuwirth, L.W., Baron-Maldonado, M., Lescop, J.M., & Paget, N., et al. (2002). Procedures for establishing defensible programmes for assessing practice performance. *Medical Education*, 36(10), 936–941.
- Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23, 4–14.
- Lunz, M.E. (1995). Statistical methods to improve decision reproducibility. In E.L. Mancall & P.G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 97–106). Evanston, IL: American Board of Medical Specialties.

- Lunz, M.E., Stahl, J.A., & Wright, B.D. (1994). Interjudge reliability and decision reproducibility. *Educational Psychological Measurement, 54*(4), 913–925.
- E.L. Mancall, P.G. Bashook (Eds.) (1995). *Assessing clinical reasoning: The oral examination and alternative methods*. Evanston, IL: American Board of Medical Specialties.
- Mancall, E.L., Bashook, P.G., & Dockery, J.L. (1996). *Computer-based examinations for board certification: Today's opportunities and tomorrow's possibilities*. Evanston, IL: American Board of Medical Specialties.
- McDermott, J., Scheiber, P.T., Juul, D., Shore, J., Tucker, G., & McCurdy, L., et al. (1991). Reliability of the part II board certification examination in psychiatry: Inter-examiner consistency. *American Journal of Psychiatry, 148*(12), 1672–1674.
- Miller, G.E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, S63–S67.
- Morris, J.A., Goplerud, E.N., & Hoge, M.A. (2004). Workforce issues in behavioral health. Institute of Medicine, Unpublished manuscript.
- Newble, D., Dauphinee, D., Macdonald, M., Mulholland, H., Dawson, B., & Page, G., et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine, 6*(3), 213–220.
- Norman, G.R., Davis, D.A., Lamb, S., Hanna, E., Caulford, P., & Kaigas, T. (1993). Competency assessment of primary care physicians as part of a peer review program. *Journal of American Medical Association, 270*(9), 1046–1051.
- Norman, G. R., Swanson, D.B., & Case, S.M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching & Learning in Medicine, 8*(4), 208–216.
- O'Sullivan, P., Cogbill, K., McClain, T., Reckase, M., & Clardy, J. (2002). Portfolios as a novel approach for residency evaluation. *Academic Psychiatry, 26*(3), 173–178.
- Page, G. (1995). Assessing reasoning and judgment. In E.L. Mancall & P.G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 19–27). Evanston, IL: American Board of Medical Specialties.
- Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194–201.
- Ramsey, P.G., Wenrich, M.D., Carline, J.D., Inui, T.S., Larson, E.B., & LoGerfo, J.P. (1993). Use of peer ratings to evaluate physician performance. *Journal of American Medical Association, 269*(13), 1655–1660.
- Regehr, G., MacRae, H.M., Reznick, R.K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance in an OSCE format examination. *Academic Medicine, 73*(9), 993–997.
- Rethans, J.J., Norcini, J.J., Baron-Maldonado, M., Blackmore, D., Jolly, B.C., & LaDuca, T., et al. (2002). The relationship between competence and performance: Implications for assessing practice performance. *Medical Education, 36*(10), 901–909.
- Roberts, C., Newble, D.I., & O'Rourke, A.J. (2002). Portfolio-based assessments in medical education: Are they valid and reliable for summative purposes? *Medical Education, 36*(10), 899–900.
- Roe, R.A. (2002). What makes a competent psychologist? *European Psychologist, 7*(3), 192–202.
- Royal College of Physicians and Surgeons of Canada. (2004). *Maintenance of certification*. Available from: <http://rcpsc.medical.org/>.
- Sabin, J.E., & Daniels, N. (2003). Strengthening the consumer voice in managed care: VII. The Georgia peer specialist program. *Psychiatric Services, 54*(4), 497–498.
- Scheiber, S.C., Kramer, T.A.M., & Adamowski, S.E. (2003). *Core competencies for psychiatric practice*. Arlington, VA: American Psychiatric Publishing.
- Schuwirth, L.W.T., Southgate, L., Page, G.G., Paget, N.S., Lescop, J.M.J., & Lew, S.R., et al. (2002). When enough is enough: A conceptual basis for fair and defensible practice performance assessment. *Medical Education, 36*(10), 925–930.
- Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park CA: Sage Publications.
- Solomon, D., Reinhart, M., Bridgham, R., Munger, B., & Starnaman, S. (1990). An assessment of an oral examination format for evaluating clinical competence in emergency medicine. *Academic Medicine, 65*(S43–S44).
- Southgate, L.J., Cox, J., David, T., Hatch, D., Howes, A., & Johnson, N., et al. (2001). The general medical council's performance procedures: Peer review of performance in the workplace. *Medical Education, 35*(Suppl 1), 9–19.
- Swanson, D.B., Norman, G.R., & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5–11.
- Taffinder, N., Sutton, C., Fishwick, R.J., McManus, I.C., & Darzi, A. (1998). Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: Results from randomised controlled studies using the MIST VR laparoscopic simulator. In J.D. Westwood, H.M. Hoffman, D. Stredney &

- S.J. Weghorst (Eds.), *Medicine meets virtual reality* (pp. 124–130). Amsterdam: IOS Press.
- Weaver, M.J., Ow, C.L., Walker, D.J., & Degenhardt, E.F. (1993). A questionnaire for patients' evaluations of their physicians' humanistic behaviors. *Journal of General Internal Medicine*, *8*, 135–139.
- Winckel, C. P., Reznick, R.K., Cohen, R., & Taylor, B. (1994). Reliability and construct validity of a structured technical skills assessment form. *American Journal of Surgery*, *167*(4), 423–427.
- World Federation for Medical Education. (1998). International standards in medical education: Assessment and accreditation of medical schools' educational programmes: A WFME position paper. *Medical Education*, *32*, 549–558.