PERFORMANCE ASSESSMENT IN MEDICAL EDUCATION
Where We've Been and Where We're Going

LISA D. HOWLEY
University of North Carolina at Charlotte

The assessment of clinical competence is becoming increasingly complex, patient centered, and student driven. Traditionally, clinical evaluation methods consisted primarily of faculty observations, oral examinations, and multiple-choice tests. Increased faculty workload, discontent with traditional methods of clinical skill assessment, and developments in the fields of psychology and education have led to the formation of new modalities, namely performance assessments. The literature pertaining to the performance assessment with standardized patients is reviewed. Based on this literature, several areas for the future direction of performance assessment are proposed, including (a) toward evidence-based locally developed assessments, (b) toward an understanding of educational outcomes and noncognitive assessment factors, and (c) toward more student-driven assessments.

Keywords: *performance assessment; medical education; standardized patient; simulated patient; clinical competence; review*

EVALUATION & THE HEALTH PROFESSIONS, Vol. 27 No. 3, September 2004 285-303
DOI: 10.1177/0163278704267044
© 2004 Sage Publications

The assessment of clinical competence is one of the most difficult tasks facing medical education. Whether the purpose is to certify a level of achievement, provide feedback to students about their clinical skills, or provide faculty with information about curriculum effectiveness, the method of assessment has a powerful effect on how and what students learn. If the assessments are inappropriate or primarily focused on basic cognitive skills, misinformation will be given back to students, and poor decisions will be made. Ultimately, inferior assessment practices will result in dissatisfied patients and compromised health care.

The field of medical education is becoming increasingly more complex. What once was considered a credible form of education and assessment now falls below our acceptable level of standards. Traditionally, clinical evaluation methods consisted primarily of faculty observations, oral examinations, and multiple-choice tests; however, as clinical faculty and house staff are faced with new demands on their time, their ability to observe students and administer oral examinations is becoming much more difficult. Consequently, several alternative or nontraditional methods for the assessment of clinical skills have been developed for local and national use. Most notably, the board overseeing the United States Medical Licensing Examination (USMLE), the agency responsible for licensing all U.S. medical doctors, recently voted to institute a new clinical skills exam—a nontraditional performance assessment (United States Medical Licensing Examination [USMLE], 2003). In this article, I provide a review of select research literature related to performance assessment and suggest three areas for future research and development in the assessment of clinical competence.

A sound assessment modality must include a clear statement of purpose, a detailed description of what is to be measured, a set of instructions for feasible administration and scoring, and guidelines for data interpretation. If intended to measure complex cognitive skills, it is reality based and taps into the high-level skills of application, analysis, synthesis, and evaluation. Finally, it also includes sufficient evidence that the scores derived from the modality are reliable and valid indicators of students' clinical competencies.

All assessments are imperfect measures of knowledge, skills, and performance. The critical question in the development and administration of an assessment is "How imperfect is it?" Fortunately, assessments do not need to be perfect to give worthwhile information about student abilities. If, however, the assessment lacks sufficient evidence to support reliable and valid interpretations, it should be used very carefully, if at all. The psychometric concepts, reliability, and validity will be referenced throughout this review. Reliability refers to the degree to which measurements provide consistent and clear information or scores (Sax, 1997). In the context of assessment, validity refers to the "degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (American Educational Research Association, 1999). It is the most important consideration in determining the quality of an assessment and refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores (Downing & Haladyna, 1997). It describes how well an assessment can be trusted to measure what it is intended to measure. As Norris and Ennis (1989) described, "An evaluation procedure is valid in a particular situation to the extent that it measures what it is supposed to measure in that situation" (p. 49). They use the term *in a particular situation* to clarify the importance of the utility and context of the instrument in establishing its validity. In other words, an assessment may provide valid information in certain contexts and not in others. These concepts of reliability and validity are considered guiding principles in relation to the assessment methods reviewed throughout the current article.

For purposes of this review, clinical competence is defined as "the ability to gather data from the patient by history and physical examination, integrate this information into a diagnostic formulation, select appropriate investigations to confirm the diagnosis, and institute efficacious management" (Norman, 1981, p. 26). Traditionally, the most prominent and heavily weighted assessment methods of clinical competence include (a) faculty observations of clinical performance with rating scales, (b) oral examinations of clinical competence, and (c) multiple-choice examinations of clinical competence. Faculty and house staff observation of students' clinical performance remains the primary evaluation method in medical education (Barzansky & Etzel, 2003).

PREVIOUS RESEARCH ON PERFORMANCE ASSESSMENT: WHERE WE'VE BEEN

Discontent with traditional methods of clinical skill assessment and developments in the fields of psychology and education were key factors in the formation of several new modalities. The most prominent and empirically based nontraditional modality is referred to as standardized patient assessment.

HISTORICAL DEVELOPMENT

Barrows and Abrahamson (1964) developed the technique of the standardized or simulated patient (SP) in the early 1960s as a tool for clinical skill instruction and assessment. During a consensus conference devoted to the use of standardized patients in medical education, Barrows (1993) described his vital role in the development of the unique modality. He was responsible for acquiring patients for the Board examinations in neurology and psychiatry and soon realized that the use of real patients was not only physically straining but also detrimental to the nature of the examination. Patients would tire and alter their responses depending on the examiner, time of day, and other factors.

Barrows also recognized the need for a more feasible teaching and assessment tool while instructing his medical students. To aid in the assessment of his neurology clerks, he coached a woman model from the art department to simulate paraplegia, bilateral Babinskis, dissociated sensory loss, and a blind eye. She was also coached to portray the emotional tone of an actual patient displaying these troubling symptoms. Following each encounter with a clerk, she would report on his or her performance. This unique standardized format caught the attention of clinical faculty and soon became a common tool in the instruction and assessment of clinical skills across all disciplines of medicine.

SPs have several advantages over the use of real patients in assessment of clinical performance. They can be trained to consistently reproduce the history, emotional tone, communicative style, and physical signs of an actual patient without placing stress on the real patient. The following additional benefits of standardized patient

assessment are described below: (a) standardization, (b) availability, and (c) cost-efficiency.

Standardized patients provide faculty with a standard assessment format. In other words, students are assessed interacting with the same patient portraying the same history, physical signs, and so on. SPs are more flexible than real patients and can be available at any time during the day and for extended periods of time. SPs can be trained to accurately and consistently record student performance and provide constructive feedback to the student, alleviating the need for direct observation by a clinical faculty member. SPs can also be trained to perform certain basic clinical procedures and, in turn, aid in the instruction of medical students. SPs allow clinical faculty the opportunity to spend less time observing long patient encounters and teaching basic clinical procedures and more time providing direct patient care and teaching more advanced skills.

Although quite flexible and variable in structure, SP assessments generally take one of the following two formats: (a) objective structured clinical examination (OSCE), or (b) the clinical practice examination (CPX). The OSCE is a limited performance assessment consisting of several brief (5 to 10 minute) stations where the student performs a very focused task, such as a knee examination, fundoscopic examination, or EKG reading (Harden & Gleeson, 1979). Conversely, the CPX is an extended performance assessment consisting of several long (15 to 50 minute) stations where the student interacts with patients in a less structured environment (Barrows, Williams, & Moy, 1987).

The format of the assessment should be driven by its purpose. If, for example, faculty members are interested in knowing how students are performing specific physical examination or radiology skills, then the OSCE format would be suitable. If, however, the faculty are interested in knowing how students are performing more complex clinical skills such as interpersonal communication, patient education, data gathering, and management, then the CPX format would be the ideal choice.

The use of standardized patients has increased dramatically, particularly over the past decade. The most recent annual survey conducted by the Liaison Committee on Medical Education (LCME) reported that 75% of U.S. medical schools are using standardized patients for evaluation within introductory skills courses: 63% are using SPs

within a comprehensive OSCE/CPX during the 3rd- or 4th year (as cited in Barzansky & Etzel, 2003). In addition, many medical licensing and specialty certification boards in the United States and Canada are either implementing or preparing to implement SP methods of assessment, including (a) the newly proposed clinical skills exam of the USMLE, (b) the Educational Commission of Foreign Medical Graduates (ECFMG), (c) the Medical Council of Canada, (d) the Royal College of Physicians and Surgeons of Canada, and (e) the Corporation of Medical Professionals of Quebec.

In their opening remarks at the first annual conference devoted to SP issues in medical education, Gliva-McConvey and Morrison (1997) reported that since the 1960s, well over 400 journal publications have cited the use of SPs. At the time, these predominantly included such issues as rater reliability, SP role consistency and accuracy, score reliability, and validity. For purposes of the current review, the literature is limited to several of the more prominent studies, which document and investigate the evidence relating to validity, reliability, scoring and standard setting methods, security, and educational outcomes of SP assessments.

PSYCHOMETRIC EVIDENCES

Validity evidence. SP-based assessments are very complex and contextual. Consequently, the research surrounding the validity of these methods is difficult and somewhat elusive. In an article describing validity in relation to SP-based assessments, Hodges (2003) argued that "our approaches to validity may themselves not be valid." He recommended that we redefine and broaden our definition of validity and not only gather quantitative evidence but also conduct "sophisticated qualitative research."

Although the concept of validity and our approaches to it are changing, previous research has focused primarily on traditional quantitative psychometric evidences (content-, criterion-, and construct-related). Relatively few studies have been conducted to assess content-related evidence of validity. A study conducted over 6 consecutive years presented a group of clinical faculty with a pre-established list of exit objectives and asked them to identify skills that they agreed should be evaluated with a performance-based examination. Thirty-two of the

36 skills were identified and 27 of those selected were covered on each of the SP assessments (Vu, Barrows, Marcy, Verhulst, & Travis, 1992).

The majority of studies investigating the validity of SP assessments have focused on criterion-related evidence (Rutala et al., 1992; Vu, Distlehorst, Verhulst, & Colliver, 1993). Rutala (1992) and Vu et al. (1993) compared the performance of residents in their 1st year of postgraduate practice with performance on an SP assessment taken during medical school. Both studies assessed the residents' performance by surveying their supervisors. Rutala et al. (1992) found moderate, significant validity coefficients between the SP assessment and resident performance ($r = .38$, $p < .01$). Because faculty ratings of student performance are known for being restricted in range and positively skewed, this moderate correlation is promising. To address the restriction and skewness problem, Vu et al. (1992) used distribution-free statistical methods (viz., frequencies and percentages) and found the SP assessment to be a "promising predictor" of resident performance. Specifically, they found the SP assessment scores better predictors of those students who received high ratings than those who received low ratings in their 1st year of residency. However, the authors concluded that these findings may be related to the low reliability of the supervisors' ratings and the low number of ratings completed for those students performing poorly on the SP assessment.

Numerous investigators have studied the relationship between existing, or concurrent, criteria and SP assessments. Swanson and Stillman (1990) summarized the findings of 10 studies and reported observed correlations ranging from 0.00 to 0.75 (faculty ratings), 0.50 to 0.77 (locally developed multiple-choice questions [MCQ]), 0.13 to 0.63 (USMLE Step I), and 0.28 to 0.60 (USMLE Step II). In a unique study, Tamblyn et al. (1994) investigated the concurrent-related validity evidence of ratings made by SPs to those of actual patients and found them to be valid predictors. However, like the Vu et al. (1992) study, the investigators found the SP ratings to lack specificity, or the ability to predict residents who were in the lowest quartile of patient ratings. The SPs reported lower satisfaction with the residents than did the actual patients. The investigators attribute this difference to the demographic discrepancies between actual and SPs. Specifically, the SPs in the current study were younger and more highly educated than the actual patients.

A few studies have been conducted to specifically investigate the evidence relating to the construct being measured in the SP assessments (Barnhart, Marcy, Colliver, & Verhulst, 1995; Newble, Hoare & Elmslie, 1981; Petrusa et al., 1987). Barnhart et al. (1995) compared the performances of 2nd- and 4th-year medical students on an SP assessment. The results expressed a sizable difference between the novice and experienced students. For example, there was a significant difference in passing rates between the 2nd-year (3%) and the 4th-year students (70%), $p = .001$. Stillman et al. (1986) found that as residents progressed through 3 years of training, their scores on an SP assessment subsequently improved. These investigators also found a significant positive relationship between performance and the prestige level of their educational institution. Other studies seem to contradict these findings (Hodges et al., 2002; Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). For example, Hodges et al. (2002) found that checklists, requiring the SP or observer to simply count the number of questions asked, were biased toward the novice student: The more experienced and efficient clinician or expert received a weaker checklist score than the novice student.

Another approach to gathering construct-related evidence is referred to as the "gold standard" or "holistic" method and involves the comparison of holistic expert ratings of students' overall performance with SP's ratings (Bardes, Colliver, Alonso, & Swartz, 1996; Croen & Moroff, 1994; Swartz, Colliver, Bardes, et al., 1997). These ratings are either dichotomous or continuous and are established by the opinion of a group or panel of experts. Bardes et al. (1996) reported validity coefficients for SP assessment scores and faculty observer ratings. They reported moderate validity correlation coefficients between SP assessment scores and faculty global ratings for two consecutive classes, 0.51 and 0.46, respectively. Both validity coefficients were statistically significant ($p < .05$). In a similar, but much larger study, Swartz et al. (1997) recruited five faculty physicians to independently observe and globally rate seven videotaped SP encounters for 44 students. The investigators found encouraging validity coefficients ranging from 0.60 to 0.70 ($p < .01$). In a study by Wilkinson and Fontaine (2002), the authors concluded that SPs themselves were able to provide reliable global or holistic ratings of students' clinical skills. These ratings were correlated with traditional

total OSCE scores ($r = .74$, $p < .001$), written ($r = .47$, $p < .001$), and in-course examinations ($r = .53$, $p < .001$) of knowledge.

Reliability evidence. The reliability of SP assessments has been the most widely investigated characteristic of this modality. Most researchers have focused on the stability of SP performance, objectivity in rating, internal consistency, and generalizability of the assessment. Several investigators have assessed the stability of SP performance across student interactions (Badger et al., 1995; Tamblyn, Klass, Schnabl, & Kopelow, 1990; Tamblyn, Klass, Schnabl, & Kopelow, 1991b; Vu, Steward, & Marcy, 1987). For example, Tamblyn et al. (1991b) reported an average accuracy rate of more than 90% in a large study including 839 SP encounters, involving 27 different cases, portrayed by 88 different SPs, trained by two trainers at two sites. Badger et al. (1995) studied the consistency of the numbers of SP-volunteered case-specific symptoms and the stability of affect and behavior over a 1-year period across doctor-patient encounters. Results from this longitudinal study revealed that multiple SPs were able to enact their roles indistinguishably from their SP counterparts. Furthermore, the investigators found evidence that performance for the majority of SPs was consistent even when intervals between simulations were as long as 3 months.

The ability of an SP to accurately, consistently, and objectively record student performance is key in SP assessments, particularly when the SP is the sole rater or recorder of student performance. For this reason many studies have been conducted to explore this important issue in SP assessment (DeChamplain, Margolis, King, & Klass, 1997; Elliot & Hickam, 1987; Finlay, Stott, & Kinnersley, 1995; Gammon, 1998; Gorter et al., 2002; Hodges, Turnbull, Cohen, Bienenstock, & Norman, 1996; Tamblyn, Klass, Schnabl, & Kopelow, 1991a; Wilkinson & Fontaine, 2002). Generally, SP ratings are either compared with the ratings of other SPs or with clinical faculty or medical educators. Two studies reporting the proportions of agreement between multiple raters show similar promising results. Specifically, 82% agreement was found between multiple SPs (Tamblyn, 1991a) and 85% agreement was found between SPs and trainers rating the same encounters (Gammon, 1998). Elliot and Hickam (1987) found that SPs with limited training were able to

reliably record 83% of the same clinical skills that were reliably evaluated by clinical faculty. DeChamplain et al. (1997) reported even higher proportions of agreement (0.88 to 0.92) in their investigation of SP accuracy as determined by the ratings of two experienced trainers.

In a review of SP assessments, Van Der Vleuten and Swanson (1990) summarized reports of reliability coefficients and test length from 13 SP-based data sets. These generalizability coefficients ranged from 0.41 for a 2-hour examination to 0.85 for a 3-hour examination. The average generalizability coefficient reported was 0.62. In addition, the authors computed the necessary hours required to achieve the recommended reliability coefficient of 0.80. These estimates ranged from 3 to 12 hours with an average test length of 7 hours.

SCORING AND STANDARD SETTING

Another important issue to consider in the development of an SP assessment is scoring and standard setting. The standard, or cut-off, score is the point above which the student must score to pass the assessment. If an SP assessment is criterion referenced, the standards required for passing the assessment must be clearly delineated.

Several standard-setting methods exist for criterion-referenced performance assessments. Berk (1986) identified and reviewed 38 methods for criterion-referenced written and performance assessments. Numerous other methods, designed specifically for SP assessments, have been cited since this comprehensive review (Clauser & Clyman, 1994; Colliver, Barnhart, Marcy, & Verhulst, 1994; Croen & Moroff, 1994; Ferrell, 1996; Morrison, McNally, Wylie, McFaul, & Thompson, 1996; Ross, Clauser, Margolis, Orr, & Klass, 1996; Travis et al., 1996). Many of these methods have been adapted from those used to set standards on multiple-choice examinations, such as the Angoff method (Angoff, 1971). This method is the most popular standard-setting method for multiple-choice examinations and has consequently been widely used with SP assessments. When the Angoff method is employed, a judge or group of judges is asked to imagine the minimally competent student and estimate that person's answers, item by item, on a given test. This minimum standard is used as a reference for assessing the groups' performance. Although popular because of the ease in implementation and computation, this

method tends to produce variable results among judges (Poggio, 1984).

Another method that is becoming increasingly popular for setting SP assessment standards is referred to as the "contrasting groups" or "expert judgment" method (Livingston & Zieky, 1982; Margolis, DeChamplain, & Klass, 1998; Ross et al., 1996). A recent review of standard-setting methods specific to SP-based examinations recommends these "examinee-centered" methods over "test-centered" methods (Boulet, DeChamplain, & McKinley, 2003). A judge or group of judges is given a sample of general student performance data (i.e., written performance or videotape performance) and asked to qualify them as masters or nonmasters. This judgmental data is then compared with the actual assessment score. The score that best discriminates between master and nonmasters is then chosen as the cut-off score for the entire group. Although several other standard-setting methods exist, further discussion is beyond the scope of this review. The reader should refer to Berk (1986), Boulet et al. (2003), Goodwin (1996), Hambleton (1995), Norcini and Shea (1997), and Margolis et al. (1998) for further descriptions of standard setting methods and criteria for implementation.

SECURITY

Because the majority of SP assessments take several days to weeks to administer, many educators have questioned the security of the case content across repeated administrations. Colliver et al. (1992) investigated the effects of repeated administration on students' working and final diagnosis scores in an SP assessments administered over a 5-year period. The authors found that the transmission of information among students had a minimal effect on a student's initial diagnosis scores and no effect on the final diagnosis scores. An additional study, conducted by Swartz, Colliver, Cohen, & Barrows (1993), deliberately requested students to "communicate as many details about the SP cases as possible to examinees tested in the second (subsequent) group" (p. S76). The authors found surprising results: The deliberate breach of test security had a minimal effect on student scores. Specifically, a nonsignificant mean increase of less than 2 percentage points was found between the two groups of students.

EDUCATIONAL OUTCOMES AND NONCOGNITIVE ASSESSMENT FACTORS

A few investigators have assessed the educational and curricular outcomes of implementing an SP assessment on students and faculty (Newble & Jaeger, 1983; Stillman, Haley, Regan, & Philbin, 1991; Ytterberg et al., 1998). Examples of these outcomes include student study time and focus, amount of faculty time spent observing student performance, and student confidence or self-efficacy with their clinical ability.

The implementation of an SP assessment has been shown to alter student study time, decreasing the amount of attention given to preparing for a multiple-choice examination and increasing the amount spent preparing for a clinical practice examination (Newble & Jaeger, 1983). Another study was conducted to assess the positive benefits of an SP assessment and found that the number of students who reported never being observed by a faculty member performing a complete history and physical examination decreased from 68% to 21% over a 4-year period (Stillman et al., 1991). These results were attributed to the implementation of an SP examination because it was the only curricular change over this period. A study about students' confidence in clinical skills revealed that participation in the SP assessment (OSCE) increased students' confidence in clinical performance (Ytterberg et al., 1998). The investigators concluded that an SP assessment that "includes challenging simulated patient situations, provides immediate feedback, and is not a high stakes exam can increase students' self-confidence in clinical skills" (p. S105).

FUTURE DIRECTIONS OF PERFORMANCE ASSESSMENT: WHERE WE'RE GOING

Based on the literature, several areas for the future direction of performance assessment have become evident. These will be addressed below and include (a) toward evidence-based locally developed assessments, (b) toward an understanding of educational outcomes and noncognitive assessment factors, and (c) toward more student-driven assessments.

TOWARD EVIDENCE-BASED LOCALLY DEVELOPED ASSESSMENTS

Despite the extensive research that has been conducted in relation to nontraditional assessments, several important areas for future research are evident, including practical development and administration issues. At her address at the 2nd annual conference of the Association of Standardized Patient Educators, Perkowski (2003) called for further research in the following areas: SP recruitment, characteristics, and training methods; performance logistics, case development, and checklist formats; and cost benefits, practice ethics, and standardized patient safety. Further support for her recommendation was provided by Gorter et al. (2000), who conducted a systematic literature review of methods used to develop checklists for use in SP-based assessments and found that "little attention" had been paid to the important process of assessment development. For another more broad review of achievements and challenges related to numerous practical considerations surrounding SP-based assessments, see Adamo (2003).

These practical issues are important considerations in relation to performance assessments and have received relatively little attention in the literature. For example, costs, staffing, resources, quality assurance training methods, and administration logistics can directly influence the validity of an assessment, and research is needed to establish best practices or standards for valid and reliable measurement practices. When administering local assessments, educators must determine whether the interpretations made from those assessments are valid. It is simply not appropriate to rely on the literature for "evidence" that a particular method is a valid indicator of performance for a locally developed measure. Validity is highly contextual and, as stated above, may provide valid information in certain contexts and not in others. Furthermore, evidence-based methods for training SPs and administering SP-based examinations are much needed.

TOWARD AN UNDERSTANDING OF EDUCATIONAL OUTCOMES AND NONCOGNITIVE ASSESSMENT FACTORS

Optimal assessments delivered at the local level have the potential to provide the students and faculty with constructive feedback about

individual and group performance. This information also has the potential to reinforce strengths and identify areas in need of remediation. Research is needed to investigate the nature and value of assessment feedback and its impact on remediation. What are the educational outcomes of completing an authentic nontraditional performance assessment? Does the assessment provide formative feedback to the learners, and if so, how effective is this feedback? Are differentiated remedial methods provided? If so, are they effective at increasing performance? Are curricular revisions made based on the assessment feedback? What are the outcomes of these revisions?

The investigation of educational outcomes and noncognitive factors related to SP assessments are important areas that may lend further credibility to the methodology. The small amount of research that has been conducted (viz., increased student self-efficacy, study time, increased faculty observation) is promising and warrants further investigation. Other potential research questions include the following: How do students' anxiety or motivation levels affect performance on a nontraditional assessment? How do students prepare or study for a nontraditional assessment? Do certain learning styles favor traditional and nontraditional assessment methods?

TOWARD MORE STUDENT-DRIVEN ASSESSMENTS

Over the years, assessments of clinical competence have become increasingly complex, patient centered, and student driven. We are experiencing a continued shift from predominantly faculty-driven and highly controlled measures to student-driven and patient-centered assessments. Student-driven assessments typically provide an unstructured environment, realistic to the natural conditions, do not limit students to lists of options, or force them to take a certain path of reasoning. For this reason, the term *authentic assessment* is often used to describe these methods (Chambers & Glassman, 1997). Faculty members merely provide a realistic context while the student directs the process. The transition from test driven to student driven is more complex than a simple change in format. Chambers and Glassman (1997) described the process as follows: "What is lost in the move from tests to authentic evaluation is faculty control over the context; what is gained is the opportunity for students to demonstrate their

ability to 'read' the real world and fashion an appropriate response out of previously learned knowledge, skills, and values" (p. 653).

Petrusa (2004) indirectly supports this shift by making several recommendations for the advancement of SP assessment and challenging us to think more broadly about these methods. She recommended several areas for advancement including expanding the traditional dyad format to include multiple-person simulations and modifying the method of measurement from checklists and rating scales to measures more capable of assessing advanced cognitive skills. These recommendations will lead to more cognitively advanced and valid measures of clinical competence.

We are moving away from limited test formats to more complex, mixed methods of authentic assessment—from faculty observation ratings supplemented with paper-and-pencil MCQ tests to SP-based performance assessment supplemented with clinical reasoning simulations. This move brings not only several unique challenges but also great educational rewards for the measurement and advancement of clinical competence.

CONCLUSIONS

As we begin this new millennium, it seems that plenty of questions remain in relation to clinical competence assessment. Although we are beginning to rely less on subjective, unreliable, and invalid methods of assessment, we still have a responsibility to continue to develop, research, and administer optimal measures. We must continue to be mindful of the inherent weaknesses with many of the traditional measures. McGuire (1995) made this point very clear in an editorial reflecting her opinions on the assessment of physician competence: "We have reduced our reliance on some of the most inherently subjective and unreliable forms of testing. . . . but we still depend far too heavily on the factual type of multiple-choice questions" (p. 740). Ironically nearly a century ago, Flexner (1910) made a similar statement in his report on the state of the medical education system in America:

> There is only one sort of licensing test that is significant, namely a test that ascertains the practical ability of the student confronting a con-

> crete case to collect all relevant data and to suggest the positive procedure applicable to the conditions disclosed. A written examination may have some incidental value; it does not touch the heart of the matter. (p. 169)

Clinical competence is an extremely complex construct and one that requires multiple, mixed, and higher order methods of assessment to support valid interpretations. Although medical students and residents are one of the most frequently tested groups in higher education, the methods of assessment are still primarily focused on low-level skills. If we expect excellence of our future physicians, we must begin to ensure competence in high-level skill areas. This begins with the use of SP-based and other more authentic clinical performance assessments. The development of optimal performance assessments, at a local or national level, is complex—requiring time, commitment, resources, and substantial efforts. However, this is the price to pay if we are to ensure clinical competence, protect the quality of patient care, and subsequently "touch the heart of the matter."

REFERENCES

Adamo, G. (2003). Simulated and standardized patients in OSCEs: Achievements and challenges 1992-2003. *Medical Teacher*, *25*(3), 262-270.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Badger, L. W., De Gruy, F., Hartman, J., Plant, M. A., Leeper, J., Ficken, R., et al. (1995). Stability of standardized patients' performance in a study of clinical decision making. *Family Medicine*, *27*, 126-131.

Bardes, C. L., Colliver, J. A., Alonso, D. R., & Swartz, M. H. (1996). Validity of standardized-patient examination scores as an indicator of faculty observer ratings. *Academic Medicine*, *71*(1), S82-S83.

Barnhart, A. J., Marcy, M. L., Colliver, J. A., & Verhulst, S. J. (1995). A comparison of second- and fourth-year medical students on a standardized patient examination of clinical competence: A construct validity study. *Teaching and Learning in Medicine*, *7*(3), 168-171.

Barrows, H. S. (1993). An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine*, *68*, 441-453.

Barrows, H. S., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Journal of Medical Education*, *39*, 802-805.

Barrows, H. S., Williams, R. G., & Moy, H. M. (1987). A comprehensive performance-based assessment of fourth-year students' clinical skills. *Journal of Medical Education*, *62*, 805-809.

Barzansky, B., & Etzel, S. I. (2003). Educational programs in US medical schools, 2002-2003. *Journal of the American Medical Association*, *290*, 1190-1196.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172.

Boulet, J. R., DeChamplain, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher*, *25*(3), 245-249.

Chambers, D. W., & Glassman, P. (1997). A primer on competency-based evaluation. *Journal of Dental Education*, *61*(8), 651-666.

Clauser, B. E., & Clyman, S. G. (1994). A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine*, *69* (10), S42-S44.

Colliver, J. A., Barnhart, A. J., Marcy, M. L., & Verhulst, S. J. (1994). Using a receiver operating characteristic (ROC) analysis to set passing standards for a standardized-patient examination of clinical competence. *Academic Medicine*, *69*(10), S37-S39.

Colliver, J. A., Travis, T. A., Robbs, R. S., Barnhart, A. J., Shirar, L. E., Vu, N. V. (1992). Test security in standardized-patient examinations: Analysis with scores on working diagnosis and final diagnosis. *Academic Medicine*, *67*(10), S7-S9.

Croen, L. G., & Moroff, S. V. (1994). Pilot-testing a holistic approach to scoring performance on standardized-patient examinations. *Academic Medicine*, *69*(4), 310-312.

DeChamplain, A. F., Margolis, M. J., King, A., & Klass, D. (1997). Standardized patients' accuracy in recording examinees' behaviors using checklists. *Academic Medicine*, *72*(10), S85-S87.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurements in Education*, *10*(1), 61-82.

Elliot, D. L., & Hickam, D. H. (1987). Evaluation of physical examination skills: reliability of faculty observers and patient instructors. *Journal of American Medical Association*, *258*(23), 3405-3408.

Ferrell, B. G. (1996). A critical elements approach to developing checklists for a clinical performance examination. *Medical Education Online*, *1*(5), 1-7.

Finlay, I. G., Stott, N. C. H., & Kinnersley, P. (1995). The assessment of communication skills in palliative medicine: A comparison of the scores of examiners and simulated patients. *Medical Education*, *29*, 424-429.

Flexner, A. (1910). *Medical education in the United States and Canada*. New York: Carnegie Foundation for the Advancement of Teaching.

Gammon, W. (1998, October). *Monitoring the quality of standardized patient ratings*. Abstract presented at the Annual Meeting of the Association of American Medical Colleges, New Orleans, LA.

Gliva-McConvey, G., & Morrison, L. (1997, September). *Professional development issues of standardized patient educators*. Invited speech at the Standardized Patient Educators Conference: Thinking Outside the Box, University of Arkansas for Medical Sciences, Little Rock, AR.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off scores. *Research in Nursing and Health*, *19*, 249-256.

Gorter, S., Rethans, J. J., Sherpbier, A., Van Der Heijde, D., Houben, H., Van Der Vleuten, C., et al. (2000). Developing case specific checklists for standardized patient- based assessments in internal medicine: A review of the literature. *Academic Medicine*, *75*(11), 1130-1137.

Gorter, S., Rethans, J. J., Van Der Heijde, D., Sherpbier, A., Houben, H., Van Der Vleuten, C., et al. (2002). Reproducibility of clinical performance assessment in practice using incognito standardized patients. *Medical Education*, *36*(9), 827-832.

Hambleton, R. K. (1995). *Setting standards on performance assessments: Promising new methods and technical issues*. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED403289)

Harden, R., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, *13*, 41-54.

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher*, *25*(3), 250-254.

Hodges, B., McNaughton, N., Regehr, G., Tiberius, R., & Hanson, M. (2002). The challenge of creating new OSCE measures to capture the characteristics of expertise. *Medical Education*, *36*(8), 742-748

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine*, *74*(10), 1129-1134.

Hodges, B., Turnbull, J., Cohen, R., Bienenstock, A., & Norman, G. (1996). Evaluating communication skills in the objective structured clinical examination format: Reliability and generalizability. *Medical Education*, *30*, 38-43.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Margolis, M. J., DeChamplain, A., & Klass, D. J. (1998). Setting examination-level standards for a performance-based assessment of physicians' clinical skills. *Academic Medicine*, *73*(10), S114-116.

McGuire, C. H. (1995). Reflections of a maverick measurement maven. *Journal of the American Medical Association*, *274*(9), 735-740.

Morrison, H., McNally, H., Wylie, C., McFaul, P., & Thompson, W. (1996). The passing score in the objective structured clinical examination. *Medical Education*, *30*, 345-348.

Newble, D., & Jaeger, K. (1983). The effects of assessments and examinations on the learning of medical students. *Medical Education*, *17*, 165-171.

Newble, D. I., Hoare, J., & Elmslie, R. G. (1981). The validity and reliability of a new examination of the clinical competence of medical students. *Medical Education*, *15*, 46-52.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, *10*(1), 39-59.

Norman, G., & Feightner, J. (1981). A comparison of behavior on simulated patients and patient management problems. *Medical Education*, *15*, 26-32.

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking: The practitioner's guide to teaching thinking series* Pacific Grove, CA: Critical Thinking Press & Software.

Perkowski, L. (2003, July). *Quality through research*. Invited speech at the 2nd Annual Conference of the Association of Standardized Patient Educators Conference: Keys to Quality, Virginia Beach, VA.

Petrusa, E. (2004). Taking standardized patient-based examinations to the next level. *Teaching and Learning in Medicine: An International Journal*, *16*(1), 98-110.

Petrusa, E., Blackwell, T. A., Rogers, L. P., Saydjari, C., Parcel, S., & Guckian, J. C. (1987). An objective measure of clinical performance. *American Journal of Medicine*, *83*, 34-42.

Poggio, J. P. (1984, April). *Practical considerations when setting test standards: A look at the process used in Kansas*. Paper presented at the 68th Annual Meeting of the American Educational Research Association, New Orleans, LA.

Ross, L. P., Clauser, B. E., Margolis, M. J., Orr, N. A., & Klass, D. J. (1996). An expert-judgment approach to setting standards for a standardized-patient examination. *Academic Medicine*, *71*(10), S4-S6.

Rutala, P. J., Fulginiti, J. V., McGeah, A. M., Leko, E. O., Koff, N. A., & Witzke, D. B. (1992). Predictive validity of a required multidisciplinary standardized-patient examination. *Academic Medicine*, *67*, S60-S62.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (3rd ed.). Belmont, CA: Wadsworth.

Stillman, P., Haley, H., Regan, M. B., & Philbin, M. M. (1991). Positive effects of a clinical performance assessment program. *Academic Medicine*, *66*(8), 481-483.

Stillman, P. L., Swanson, D. B., Smee, S., Stillman, A. E., Ebert, T. H., Emmel, V. S., et al. (1986). Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine*, *105*(5), 762-771.

Swanson, D. B., & Stillman, P. L. (1990, March). Use of standardized patients for teaching and assessing clinical skills. *Evaluation & the Health Professions*, *13*(1), 79-103.

Swartz, M. H., Colliver, J., Bardes, C. L., Charon, R., Fried, E. D., & Moroff, S. (1997). Validating the standardized-patient assessment administered to medical students in the New York City Consortium. *Academic Medicine*, *72*(7), 619-626.

Swartz, M. H., Colliver, J. A., Cohen, D. S., & Barrows, H. S. (1993). The effect of deliberate, excessive violations of test security on performance on a standardized-patient examination. *Academic Medicine*, *68*(10), S76-S78.

Tamblyn, R., Abrahamowicz, M., Schnarch, B., Colliver, J. A., Benaroya, S., & Snell, L. (1994). Can standardized patients predict real-patient satisfaction with the doctor-patient relationship? *Teaching and Learning in Medicine*, *6*(1), 36-44.

Tamblyn, R. M., Klass, D. K., Schnabl, G. K., & Kopelow, M. L. (1990). Factors associated with the accuracy of standardized patient presentation. *Academic Medicine*, *65*(9), S55-S56.

Tamblyn, R. M., Klass, D. J., Schnabl, G. K., & Kopelow, M. L. (1991a). The accuracy of standardized patient presentation. *Medical Education*, *25*, 100-109.

Tamblyn, R. M., Klass, D. J., Schnabl, G. K., & Kopelow, M. L. (1991b). Sources of unreliability and bias in standardized-patient rating. *Teaching and Learning Medicine*, *3*(2), 74-85.

Travis, T. A., Colliver, J. A., Robbs, R. S., Barnhart, A. J., Barrows, H. S., Giannone, L., et al. (1996). Validity of a simple approach to scoring and standard setting for standardized-patient cases in an examination of clinical competence. *Academic Medicine*, *71* (1), S84-S86.

United States Medical Licensing Examination. (2003). *Americans overwhelmingly support new medical license test: Field trials show fairness, reliability of test*. Retrieved August 8, 2003, from www.usmle.org/news/cse/newsrelease2503.htm

Van Der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, *2*(2), 58-76.

Vu, N. V., Barrows, H. S., Marcy, M. L., Verhulst, S. J., & Travis, T. (1992). Six years of comprehensive clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Academic Medicine*, *67*, 42-45.

Vu, N. V., Distlehorst, L. H., Verhulst, S. J., & Colliver, J. A. (1993). Clinical performance-based test sensitivity and specificity in predicting first year residency performance. *Academic Medicine*, *68*(2), S41-S45.

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

Wilkinson, T. J., & Fontaine, S. (2002). Patients' global ratings of student competence. Unreliable contamination or gold standard? *Medical Education*, *36*, 1117-1121.

Ytterberg, S. R., Harris, I. B., Allen, S. S., Anderson, D. C., Kofron, P. M., Kvasnicka, J. H., et al. (1998). Clinical confidence and skills of medical students: Use of an OSCE to enhance confidence in clinical skills. *Academic Medicine*, *73*(10), S103-S105.