Use of Standardized Patients in Clinical Performance Assessments: Recent Developments and Measurement Findings

Vu, Nu Viet; Barrows, Howard S.

How to cite

VU, Nu Viet, BARROWS, Howard S. Use of Standardized Patients in Clinical Performance Assessments: Recent Developments and Measurement Findings. In: Educational researcher, 1994, vol. 23, n° 3, p. 23–30. doi: 10.3102/0013189X023003023

This publication URL: https://archive-ouverte.unige.ch/unige:26094

Publication DOI: 10.3102/0013189X023003023

Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings

Nu Viet Vu and Howard S. Barrows

EDUCATIONAL RESEARCHER 1994 23: 23
DOI: 10.3102/0013189X023003023

Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings

Nu VIET VU HOWARD S. BARROWS

This article reviews recent developments and measurement findings on the use of live patient simulations or "standardized patients" in performance examinations to assess the competence of medical professionals. The results of large-scale standardized patient-based performance assessments are presented and discussed in terms of their feasibility, reliability, validity, and implications for assessing competence in other professions.

Educational Researcher, Vol. 23, No. 3, pp. 23-30.

Similar to recent trends in mathematics, science, English, architecture, and law, the growing movement to assess performance in medicine indicates that competence can no longer be validly assessed solely by writ ten examinations such as multiple-choice tests (Mehrens, 1992); The performance capability of physicians is found to be more relevantly assessed directly in the context of situations or problems commonly encountered in medical practice. Typically, performance assessments in medicine consist of having the examinees encounter a series of patient problems, in which they are expected to assess, evaluate, and resolve the issues or problems brought in by the patient. Unlike multiple-choice tests, performance tests allow the examinees to be observed and assessed directly on their ability (a) to perform various clinical skills (e.g., interviewing and examining the patient) and/or technical procedures (e.g., inserting an IV tube) and (b) to communicate, relate, counsel, and educate the patient(s).

Since the main objective in the training of physicians is to ensure that they will deliver competent, quality care when encountering various patient problems, different formats of performance assessments have been developed to evaluate clinical competence (McGuire, 1983; Morgan & Irby, 1978; Stillman & Gillers, 1986; Vu, l979). The types of performance assessments that have commonly been used until now are (a) the direct observation and evaluation of students' performance with a real patient through the use of a rating scale, and (b) the indirect evaluation of students' performance with a real patient through their oral or written reports. Ideally, students should be repeatedly evaluated in real situations with real patients. Unfortunately, these performance assessments have both practical and measurement problems. These include the unfairness to the real patients, who have to endure being worked up by a number of students; the lack of standardization because different patients are presented to different students; the lack of objectivity in the rating and scoring, which may not be fully understood by the faculty rater; and the question of authenticity of the students' reports since the attending physicians and nurses have already recorded most patients' information on their charts. Finally, and most problematic, is the availability of a reasonable number of observations for each student to be assessed in a valid and reliable manner. In a survey of medical schools and residency programs, it was found that students' and residents' performances with patients are rarely observed, and for students and residents who are observed, observation rarely takes place as often as three times during their clinical training (Stillman, Regan, & Swanson, 1987). This problem is common to many fields in which performance assessments are based on direct observations, such as the assessment of teachers' skills in real teaching situations (Sweeney & Manatt, 1986) or of students' speaking skills in "spontaneous assessments" (Stiggins & Bridgeford, 1986).

In an attempt to derive a more standardized, objective, and less time-consuming performance assessment, various types of patient simulations have been developed to replace the real patients. While earlier reviews dealt with written and computerized patient simulations (McGuire, 1983; Norman, Muzzin, Williams, & Swanson, 1985; Swanson, Norcini, & Grosso, 1987; Vu, 1979), the purpose of this article is to review the recent use of live patient simulations, referred to as simulated or standardized patients, in clinical performance assessments. This article updates and expands an earlier report by van der Vleuten and Swanson (1990) by reviewing findings obtained in the last 5 years on the validity and reliability of standardized patient-based performance assessments. In addition, it reviews important issues not previously covered on the use of live standardized simulations in performance tests such as test feasibility, portability, fidelity, and security. Last, the article discusses how the simulations can be applied in educational and professional assessments. One of the main concerns in the use of performance examinations is their feasibility when used on a large scale. For this reason, only studies that administer the examination to a large number of examinees are reviewed here.

Nu VIET VU *is professor at the Southern Illinois University School of Medicine, Department of Medical Education, P. O. Box 19230, Springfield, IL 62794. She specializes in testing and measurement in the areas of clinical problem solving and clinical performance.* HOWARD S. BARROWS IS *chairman of the Southern Illinois University School of Medicine, Department of Medical Education, P. O. Box 19230, Springfield, IL 62794. He specializes in problem-based learning and performance-based assessment.*

Standardized Patients

Reviews on the use of written and computerized simulations to assess clinical competence (McGuire, 1983; Norman et al., 1985; Swanson et al., 1987) suggested that although they provide a reasonably realistic assessment, cueing from the options provided in the simulations not only does not simulate well the free inquiry in a real patient-doctor encounter but also affects examinees' test performance in that they are found to gather more data than they would in an uncued encounter (Norman & Feightner, 1981; Page & Fielding, 1980). In addition, these simulations do not challenge examinees' abilities to observe patients, and to apply interpersonal and communication skills, and psychomotor skills of the physical examination.

Since the data-gathering process in patient-doctor encounters is a free-inquiry process, the assessment would be more accurate if inquiry could be carried out in an uncued and open-ended format. With the development of the live simulated patient (Barrows, 1987; Barrows & Abrahamson, 1964), referred to now as the standardized patient (SP), it has become possible to design an uncued, open-ended, standardized, and more objective assessment of the examinees' skills in gathering history and physical examination data from the patient. The SP is a real or simulated patient carefully coached to present a patient problem accurately and in a standardized manner for all examinees. The technology of the live and standardized patient simulation, although originally developed for the assessment of competence in medicine, can easily be adapted to assess competence in any profession in which the objectives are to assess an individual's interpersonal and communication skills as well as his or her professional and technical performance. The technique of training a standardized patient can be readily applied in educational and professional testing to train a standardized client, physician, judge, teacher, student, parent, customer, and so forth. For example, simulated-standardized students can be used to observe and assess a teacher's skills in diagnosing and counseling students in academic difficulty, in handling various levels of student questions on a topic, or in teaching and conducting a small group tutorial.

SPs are most commonly used in the multiple-station format where the examinees are presented with a series of simulated patient problems. For each patient problem, there are two stations. The first is an *encounter station*, where the examinees encounter the SP and are assessed on their ability to perform various clinical skills (e.g., history taking, physical examination, and technical procedures), to communicate verbally, to employ interpersonal skills, and to relate professionally to the patient. After the encounter station, the examinees often have a written, computerized, or oral *test station* designed to assess their problem-solving and decision-making skills within the context of the patient problem.

Test Feasibility, Portability, Fidelity, and Security

The perceived complexity of using large-scale SP-based performance assessments often raises concerns about their feasibility. Several studies have demonstrated their feasibility when examinations consisting of 3 to 25 stations, with station length varying between 5 and 40 minutes, were administered to groups of 40 to 260 candidates (Cohen et al., 1987, 1988; Grand'Maison, Lescop, & Brailovsky, 1992; Klass et al., 1987; Newble & Swanson, 1988; Petrusa et al., 1987; Reznick et al., 1992; Stillman et al., 1986a, 1987; Tamblyn, Klass, Schnabl, & Kopelow, 1991a, 1991b; Vu, Barrows, et al., 1992; Williams et al., 1987). Depending on the station length and the number of stations and candidates to be tested on the examination, the administration of the examination can vary from half a day (Petrusa et al., 1987) to 3 weeks (Vu, Barrows, et al., 1992). In the latter situation, examinees are put into groups that are tested on different days.

Overall, SP-based performance assessments have been shown to be transportable across testing centers as no significant differences were found in the exam scores of candidates taking the same test administered at different sites (Grand'Maison et al., 1992; Klass et al., 1987; Reznick et al., 1992; Sutnick, 1991) or when translated into different languages (Reznick et al., 1992). It was also found that with careful training there were no significant differences in the SPs' portrayal of the same case across testing centers for the majority of cases in the assessment (Reznick et al., 1992; Tamblyn et al., 1991a). It was found that although minor inaccuracies in the simulations could affect students' performance or pass-fail rate on individual cases (Dawson-Saunders, Verhulst, Marcy, & Steward, 1987; Tamblyn et al., 1991a), such inaccuracies did not seem to greatly affect the examinees' overall exam mean scores and overall exam failure rates (Colliver, Robbs, & Vu, 1991; Reznick et al., 1992) or the reliability of the overall exam mean scores (Colliver, Morrison, Markwell, Verhulst, & Steward, 1990; Swanson & Norcini, 1989).

Regarding the authenticity of SPs, it was shown that SPs were usually not detected by family physicians when the SPs were sent to their practice (Burri, McCaughan, & Barrows, 1976; Owen & Winkler, 1974; Rethans, Sturmans, Drop, & van der VIeuten, 1991), and when the SPs were detected, it was at a negligible rate of 2% (Neufeld, Woodward, & Norman, 1983) to 4% (Rethans, Drop, Sturmans, & van der VIeuten, 1991; Rethans & van Boven, 1987). It was also found that there was no difference in the physicians' performances when they worked up real patients as opposed to SPs (Norman & Tugwell, 1982), and that their performances with SPs' simulations reflected their actual performances more than written simulations did (Rethans & van Boven, 1987).

One common concern regarding performance assessment is the security of the test content and the test itself given the limited number of tasks on a typical test and the numerous groups of examinees taking the test at different times, often over several days (Mehrens, 1992). Preliminary results have shown that there was little obvious information sharing, and when there was, the information shared did not seem to affect the examinees' scores significantly and in any consistent manner across different patient cases and student groups (Colliver, Barrows et al., 1991; Rutala, Witzke, Leko, Fulginiti, & Taylor, 1991; Stillman, Haley, et al., 1992; Vu, Barrows, et al., 1992; Williams, Lloyd, & Simonton, 1992). These results were observed both when the examination did not count toward examinees' promotion and when it did.

Test Reliability

With the use of SPs in clinical performance assessments, many issues of reliability need to be addressed. For example, it is important to determine the accuracy and reliability of the SPs' simulations as well as of their recording of

examinees' performances on checklists. In general, it was found that SP-based performance assessments have moderate score reliability, which is due in great part to the variability of examinees' performance from task to task. It was also found that with good training, the SPs can be accurate and consistent in the essential features of their simulations as well as in recording the students' performance on checklists for faculty's evaluation.

Reliability of Test Scores and Pass-Fail Decisions

The reliability of test scores and pass-fail decisions of most large-scale SP-based performance assessments has been estimated using generalizability (Brennan, 1983) instead of classical test theory. Until now, the test scores of most standardized performance-based assessments, irrespective of format, skills assessed, and total testing time, have been estimated to have generalizability coefficients varying between .41 and .85, with most of the coefficients in the moderate range, around .50 to .60 (Grand'Maison et al., 1992; Reznick et al., 1992; van der Vleuten & Swanson, 1990; Vu, Barrows, et al., 1992). It was found that the moderate size of the reliabilities often resulted from the problem of content specificity, that is, the variability of examinees' performances across different tasks or patient cases, even when the cases were derived from the same specialty (Norman, 1985) or when the SPs presented with the same diagnosis but different presenting complaints, or with the same presenting complaint but different diagnoses (Norman, Feightner, Tugwell, Muzzin, & Guyatt, 1983). In general, the moderate reliability and content specificity are typical not only of assessments using live SP simulations but also of assessments using written and computerized simulations. In addition, they are also typical of performance assessments in the military (Shavelson, Mayberry, Li, & Webb, 1990), direct writing (Breland, Camp, Jones, Morris, & Rock, 1987; Hieronymus & Hoover, 1987), and science (Shavelson, Baxter, & Pine, 1992) as well as in the health professions. On the one hand, these results seem to suggest that further test developments are needed in order to improve the test generalizability, but on the other hand they seem to suggest that performance or competence in a field may be a relatively specific and not a generalizable measure and that higher reliability for performance tests may not be easily attained.

When standardized performance assessments are used to classify examinees as masters or nonmasters, the reproducibility of test scores would be less critical than the reproducibility of pass-fail decisions. Except for a few preliminary reports (Colliver et al., 1993; Grand'Maison et al., 1992; Reznick et al., 1992; Rothman, Cohen, Dirks, &Ross, 1990; Vu, Barrows, et al., 1992) on the use of experts' judgment in setting performance standards, research findings in this area are still limited. Overall, the problems encountered with establishing reliable pass-fail decisions in performance assessments are similar to those encountered in multiple-choice tests, and much work is still needed on how to derive and set valid pass-fail standards for complex and multiskill performances, and to determine which method of standard setting is most valid, practical, and cost-effective.

Accuracy and Consistency of Simulations

With the use of SPs, one element that needs to be assessed is the accuracy and consistency (reliability) with which the SPs simulate a problem. The accuracy of a simulation is defined as the proportion of essential features that the SP presents correctly in each patient-student encounter (Tamblyn et al., 1991a). These features are the ones presented in the history and physical examination and those related to the patient's affect. Studies have shown that with careful training techniques and experienced SP trainers, the SPs can achieve average accuracy of between 90.2% and 93.4% (Tamblyn et al., 1991a) and that they can maintain their accuracy in most of the case simulations on an examination (Reznick et al., 1992; Tamblyn et al., 1991a) and over the course of a 1-day examination (Vu, Steward, & Marcy, 1987).

Comparison of SPs' Recordings and Expert Observers' Ratings of Performance

In clinical performance assessments, examinees' performance has been evaluated in two ways. One approach is to have the examinees' performance directly observed and rated by expert raters or faculty on detailed behavior checklists. It was found that with careful rater training and detailed behavior checklists, the interrater reliability (intraclass correlation) in clinical performance assessments was relatively good, ranging from .68 to .79 (Newble & Swanson, 1988; van der Vleuten, van Luijk, & Swanson, 1988). Overall, these reliabilities are comparable to those found in direct writing (Dunbar, Koretz, & Hoover, 1991; Hieronymus & Hoover, 1987) and science assessments (Shavelson, Baxter, &,Pine, 1992). Since there is a high interrater agreement, van der Vleuten and Swanson (1990) have further illustrated and recommended that in order to derive more reproducible test scores in testing situations where the number of raters are limited, it would be more effective to increase the number of cases on the exam and decrease the number of raters assigned to each case than to do it the other way around.

In order to reduce the cost of using faculty as observer-raters, some testing centers used a second method of evaluation. They have trained the SPs to record on checklists (but not to evaluate) examinees' performance on selected items of history, physical examination, patient education, and/or counseling. The examinees' recorded performance is then evaluated and scored based on faculty's predetermined protocols and criteria. Research has shown that with training, laypersons' accuracy in recording on a checklist approaches the accuracy of the teaching staff (Elliott & Hickam, 1987; van der Vleuten, van Luijk, Ballegooijen, & Swanson, 1989) and that there is a high percentage of agreement (80% to 100%) between SPs' recordings and those of faculty and nonfaculty observers (Norman et al., 1985; Rethans & van Boven, 1987; Tamblyn et al., 1991b; Vu, Marcy, et al., 1992; Williams et al., 1987). In addition, the SPs' recordings were found to be highly consistent (82% to 85%) in test-retest situations (Rethans & van Boven, 1987; Tamblyn et al., 1991b) as well as over the course of a 1-day or a 3-week examination (Vu, Marcy, et al., 1992).

One common task of the SPs in performance assessments is to rate examinees' communications skills, interpersonal skills, and professional behavior and service from the patient's point of view. In general, the SPs' ratings were found to be reliable or generalizable (.69 to .83) when they were obtained across several problems and hence different SPs (Vu, Marcy, Verhulst, & Barrows, 1990), and the behavioral characteristics that the SPs use to derive their ratings of examinees' performances were found to be similar to those

used by real patients (Vu, Marcy, et al., 1990; Webster, 1989). Interestingly, it has also been found that when SPs are trained to evaluate examinees' English proficiency, they can provide evaluations as good as, and more reliable than, those of professional raters at Educational Testing Service (Friedman et al., 1991).

Test Validity

Validation studies of SP-based performance assessments have been conducted using both criteria defined in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) and those recently described by measurement specialists as more relevant to performance assessments (Kane, 1987; Linn, Baker, & Dunbar, 1991; Moss, 1992). Recent validity studies of performance tests suggest that the "key features" approach is a promising way to define the test content domain, that the tests assess a cognitive and a noncognitive component of clinical performance, that they provide differential performances for examinees at different training levels, and that the tests' concurrent and predictive validities remain difficult to assess in the absence of a "gold standard" of clinical competence. Attempts to evaluate aspects of validity considered more relevant to performance assessments indicate that examinees regarded the tests as fair and meaningful, that the tests assess and challenge examinees' clinical and cognitive skills, and that there is no evidence that the SPs introduce test bias into the evaluation process.

Test Content Coverage

One important validity issue in standardized performance assessments is the adequacy and representativeness of the test content since it is impractical and impossible to assess a large domain of cases and skills with this type of assessment. In most studies the issue of sampling cases and skills within and across cases has not been adequately described, and only a few studies (Grand'Maison et al., 1992; Petrusa et al., 1987; Vu, Barrows, et al., 1992) have reported careful attempts at defining and selecting from the domain. Given that most studies reported moderate test score reliability when tests included as many cases as could be practically managed and administered at one time, the question is whether there are alternative ways of defimning the domain and selecting and designing cases so that valid and reliable test scores could be derived from a limited domain. One attempt has been made by defining a domain with "key features" of a case. Although the key features have been developed for assessments using written simulations, they may be used to select and design cases for assessments using SPs. The key features of a case include those elements identified as critical in the resolution of the case, those most likely to lead to errors in the resolution of the case, and those most difficult in the identification and resolution of the case (Bordage & Page, 1987; Page, Bordage, Harasym, Bowmer, & Swanson, 1990). It is hoped that with this domain definition, each case will be more focused and shorter and, hence, that more cases can be sampled on a test. Preliminary results have suggested that the key features approach has validity: The key features pre-identified for a set of cases were found to be the same ones that were independently identified by a group of experts when they reviewed the cases (Brailovsky, Bordage, Carretier, & Page, 1922). No results are available yet on the efficiency and effectiveness of this type of domain definition on overall test validity and reliability.

Construct Validity

Two types of construct validity were conducted. Because standardized clinical performance assessments are designed to assess examinees' clinical skills in different situational tasks, attempts have been made to derive a parsimonious, yet informative and valid, way of reporting students' performance. Factor analyses of the skills assessed across different tasks have consistently identified two separate, independent factors: one cognitive and one noncognitive. The cognitive factor includes skills such as gathering data, formulating working and final diagnoses, test selection, test interpretation, and patient management. The noncognitive factor include communication and interpersonal skills and professional behavior and service. Although the structure of the noncognitive factor was found to be unified and stable across examinations, the structure of the cognitive factor was found to vary and to be less stable across examinations (Hassard, Campbell, Klass, Kopelow, & Schnabl, 1990; Vu, Colliver, & Verhulst, 1992). These findings, which are consistent with those from other formats of assessment (e.g., residency supervisor ratings), suggested that the cognitive and noncognitive factors are two separate, independent aspects of clinical performance and should be reported as such (Verhulst, Colliver, Paiva, & Williams, 1986).

Another type of construct validity study consists of group differential performances. When examinees at various levels of training were compared on their clinical skill performances, it was found that they performed differently on these skills, that they had different pass-fail and odds-ratio failure rates (Barnhart, Marcy, Colliver, & Verhulst, 1992; Klass et al., 1990; Rothman et al., 1990; Stillman et al., 1986a), and that the observed differences were more accentuated with skills that require greater knowledge, training, and interpretation (e.g., diagnostic and management skills as opposed to history taking and physical examination).

Criterion Validity

Since clinical performance tests are developed to complement the shortcomings of existing ones, attempts at assessing test criterion validity (e.g., concurrent validity) by correlating scores on clinical performance tests with existing measures of clinical knowledge (e.g., standardized objective multiple-choice tests) and clinical competence (e.g., faculty ratings) have not provided useful information except for the conclusion that they are moderately correlated with one another (Petrusa et al., 1987; Reznick et al., 1992; Stillman et al., 1987; Vu, Barrows, et al., 1992). The correlations among the measures suggest that they may be complementary to one another and assess different components of clinical competence. This explanation itself is difficult to verify since there is no "gold standard of competence" against which the validity of performance measures can be adequately and ultimately assessed. Until now, the predictive value of performance measures were assessed mainly with available (e.g., residency performance ratings) instead of ultimate performance criteria (Rothman et al., 1990; Vu, Distlehorst, Verhulst, & Colliver, 1993). Again, the absence of a gold standard is not unique to the health professions but also applies to other professions. Further assessment

of the validity of performance tests, especially their concurrent validity, would only yield limited useful information. Efforts should be devoted instead to issues that are more relevant to performance assessments, such as assessing their impact on faculty's teaching and students' learning. Such studies not only are feasible (Newble & Jaeger, 1983; Stillman, Haley, Regan, & Philbin, 1991), but would provide more informative data and validate the assumption that this type of testing influences and redirects faculty's teaching and students' learning, making both more relevant to actual practice.

Test Fairness and Cognitive Complexity

Because performance tests assess examinees' ability to use their skills to "do what is needed" over a sample of professional and realistic patient cases, the validity of the cases and skills on the test needs to be evaluated in terms of their fairness, meaningfulness, and cognitive complexity (Linn et al., 1991).

It was found in one survey that, although the examinees indicated they had not previously seen or directly worked up several cases on the test, most of them had read about the cases (Vu, Barrows, et al., 1992). It was also found that both examinees and examiners regarded SP-based performance assessments as relevant and meaningful, as more appropriate and fairer measures of clinical competence than traditional clinical examinations (Newble & Jaeger, 1983), and as challenging their clinical and cognitive skills (Shirar, Vu, Colliver, & Barrows, 1992).

In using SPs in performance assessments, it is important to determine whether the SPs introduce any bias into the evaluation process. It has been found that SPs' gender and age have no main effect and do not interact with examinees' gender to affect their performances (Colliver, Marcy, Travis, & Robbs, 1991; Rutala, Witzke, Leko, & Fulginiti, 1991; Stillman, Regan, Swanson, & Haley, 1992; Vu , Marcy, et al., 1990). No evidence is yet available on whether the race of SPs and examinees has a main effect or whether these factors interact with each other to affect examinees' performances.

Test Scoring Validity

Few findings are yet available on how to derive a valid scoring key and scoring process for performance assessments. Existing findings showed that the use of different weighting of case test items or options, depending on their importance and appropriateness to the assessed task, has little impact on the resulting scores (Stillman et al., 1986b). In addition, empirical or expert-derived scoring, such as aggregate scoring, where the score of an item is proportional to the degree of agreement between experts (Norman, 1985), is found to correlate with other type of scoring and neither affects nor improves the validity of test scores (Webster, Shea, Norcini, Grosso, & Swanson, 1988). Finally, preliminary results showed that a scoring that takes into account both the examinees' performance outcomes and their underlying reasoning seems to be more discriminating than one that relies solely on the outcomes (Vu, Lee, & Steward, 1990).

With regard to the checklist scores obtained from the SP-examinee encounter, it was found that for some cases these scores concur more with experts' ratings of examinees' performances than for others (MacRae, 1991). For example, there was less concurrence in those cases where the items performed and the order in which they should be performed are not recognized as critical and standard by all experts (e.g., working up a patient complaining of headache) than in cases where all experts agree (e.g., emergency management of a patient with an abdominal pain resulting from a car accident). Since clinical performance has been shown to be content specific, it is not surprising to find that experts' derived scoring would also be content specific.

Overall, unlike multiple-choice testing, where the exam score reflects the number of correct answers, the numerical score on a performance assessment is difficult to interpret since it represents several skills, and examinees can obtain the same numerical score through different actions or patterns of actions. Reporting students' performance in terms of a skill performance profile has not been viable until now due to the low reliability of the skill scores (Colliver, Vu, Markwell, & Verhulst, 1990). Much work still remains to be done to determine how to derive valid performance scores and how they can be meaningfully reported. With the low reliability often observed with performance assessment scores, it may be necessary not only to increase the number of cases on the examination but also to increase the signal-to-noise ratio at the case level to derive more valid and reliable scores.

Conclusions

Use of live SP technology in large-scale performance assessments in the health professions has demonstrated that, at least at the institutional level, such assessments not only are feasible but also can be standardized and scored in an objective manner. When appropriately trained, SPs have been shown to provide consistent and highly accurate simulations and recordings of examinees' performances. Given these characteristics, the SP technology is useful in the assessment of performance not only in the health professions but also in any professions where it is important to assess examinees' ability to determine *on their own* the necessary tasks to be performed in a situation, to carry out those tasks correctly, and to interact and relate effectively with the individual(s) encountered in the situation. As described, the SP technology can be used in educational and professional assessments to develop standardized simulations of teacher, student, parents, or client. For example, to evaluate teachers' tutoring skills, a small group of standardized students can be trained to simulate different types of learners often encountered in real tutorial groups. With the use of standardized students, the teachers' skills can be assessed more effectively because the teachers are directly observed interacting with the students and in situations regarded as representative and common in teaching practice. The teachers are also assessed in a more objective and standardized manner because their performances are evaluated against predefined and accepted scoring criteria.

Preliminary results indicate that standardized patient-based performance assessments demonstrate reasonable evidence of validity, whether the evidence is derived from classical criteria such as test content, construct, and criterion validity, or from new validity criteria suggested by measurement specialists for performance assessments, such as test fairness, meaningfulness, and cognitive complexity. Evidence on reliability indicates that although the reliability of exam scores is not greatly affected by the variability of SPs and raters, it is greatly affected by the variability of ex-

aminees' performances across tasks. With moderate exam score reliability, specific consideration should be given to situations in which the scores are used for promotion or certification, such as the reliability or generalizability of pass-fail decisions, and how performance scores can be used in combination with other types of scores to obtain more reliable estimates of examinees' performances and pass-fail decisions.

Last, but not least, is the issue of cost of performance assessments. It goes without saying that a valid measurement of performance would entail higher costs not only in development but also in administration and scoring. Until now, decisions not to use performance assessments have been based on the high direct costs of such examinations, and no consideration has been given to the cost-benefit ratio that would result if such examinations were used. This means that the cost of using such examinations should be considered in relation to the cost of training the examinees, detecting performance errors and preventing future costly ones, and deriving specific and useful performance feed back for faculty and students.

As with any other type of testing, additional evidence needs to be obtained concerning the feasibility, reliability, and validity of performance assessments to improve them further. Present results, as reviewed here, indicate that large-scale performance assessment using SP technology is feasible and provides a relatively efficient, valid, and moderately reliable method of assessing professional competence.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

Barnhart, A., Marcy, M. L., Colliver, J. A„ Verhulst, S.]. (1992, April). *A comparison of second- and fourth-year medical students on a standardized-patient (SP) examination of clinical competence: A construct validity study.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Barrows, H. S. (1987). *Simulated (standardized) patients and other human simulations: A comprehensive guide to their training and use in teaching and evaluation.* Chapel Hill, NC: Health Sciences Consortium.

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.

Bordage, G., & Page, G. (1987). An alternative approach to PMPs: The "Key Features" Concept. In I. R. Hart & R. M. Harden (Eds.), *Further development in assessing clinical competence.* Montreal: Can-Heal.

Brailovsky, C, Bordage, G., Carretier, H., &Page, G. (1992). *Content validity of the key features approach of the Medical Council of Canada's exam.* Abstract presented at the annual meeting of the Association of American Medical Colleges Research in Medical Education, New Orleans.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. D., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph N. 11). New York: College Entrance Examination Board.

Brennan, R. (1983). *Elements of generalizability theory.* Iowa City: American College Testing Program.

Burri, A., McCaughan, K., & Barrows, H. (1976). Using stimulated patients to evaluate practicing physicians in a community. *Proceedings of the 15th Conference on Research in Medical Education,* 295-299.

Cohen, R., Rothman, A. I., Ross, J., Domovitch, E., Jamieson, C, Jewett, M., Keystone, J., Kulesha, D., Maclnnes, A., McLeary, P., Ouchterlony, D., Petrusa, E., Poldre, P., Robb, K., Rossi, M., Sarin, M., Shier, R. M., Schwartz, M. (1988). A comprehensive assessment of graduates of foreign medical schools. *Annals of the Royal College of Physicians and Surgeons of Canada, 21,* 505-509.

Cohen, R., Rothman, A. I., Ross, J., Maclnnes, A., Domovitch, E., Jamieson, C, Jewett, M., Keystone, J., Kulesha, D., McLeary, P., Ouchterlony, D., Poldre, P., Robb, K., Rossi, M., Sarin, M., Schwartz, M., Sherman, R., Shier, M. (1987). Comprehensive assessment of clinical performance. In I. R. Hart & R. M. Harden (Eds.), *Further developments in assessing clinical competence.* Montreal: Can-Heal.

Colliver, J. A., Barrows, H. S., Vu, N. V., Verhulst, S J., Mast, T. A., & Travis, T. A. (1991). Test security in examination using standardized patient cases for five classes of senior medical students. *Academic Medicine, 66,* 279-282.

Colliver, J. A., Marcy, M. L., Travis, T. A., & Robbs, R. S. (1991). The interaction of student gender and standardized patient gender on a performance-based examination of clinical competence. *Academic Medicine, 66,* 531-533.

Colliver, J. A., Morrison, L. J., Markwell, S. J., Verhulst, S. J., & Steward D. E. (1990). Three studies of the effect of multiple standardized patients on intercase reliability of five standardized-patient examinations. *Teaching and Learning in Medicine, 2,* 237-245.

Colliver, J. A., Robbs, R. S., & Vu, N. V. (1991). Effects of using two or more standardized patients to simulate the same case on case means and case failure rates. *Academic Medicine, 66,* 200-202.

Colliver, J. A., Vu, N. V., Markwell, S. J., & Verhulst, S. J. (1990). Psychometric properties of the clinical competencies assessed with standardized patient cases. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier & R. P. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence* (pp. 571-577). Groningen, Netherlands: Boekwerk.

Colliver, J. A., Vu, N. V., Robbs, R. S., Verhulst, S. J., Travis, T. A., & Barrows, H. S. (1993). False-negative and false- positive rates resulting from measurement error for a mastery examination of clinical competence based on standardized patient cases. *Teaching and Learning in Medicine, 4,* 238-242.

Dawson-Saunders, B., Verhulst, S., Marcy, M., & Steward, D. (1987). Variability in standardized patients and its effect on student performance. In I. R. Hart & R. M. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 451-456). Montreal: Can-Heal.

Dunbar, S. D., Koretz, D. M., & Hoover H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4,* 289-303.

Elliott, D. L., & Hickam, D. H. (1987). Evaluation of physical examination skills: Reliability of faculty observer and patient instructors. *Journal of the American Medical Association, 258,* 3405-3408.

Friedman, M., Sutnick, A. I., Stillman, P. L., Norcini, J. J., Anderson, S. M., Williams, R. G., Henning, G., & Reeves, M. J. (1991). The use of standardized patients to evaluate the spoken- English proficiency of foreign medical graduates. *Academic Medicine, 66,* 561-563.

Grand'Maison, P., Lescop, J., & Brailovsky, C. (1992). Large scale use of an objective structured clinical examination for licensing family physicians. *Canadian Medical Association, 1146,* 1735-1740.

Hassard, T. M., Campbell, C, Klass, D., Kopelow, M., & Schnabl, G. (1990). An examination of the factor structure of clinical competence. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, & R. P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 353-356). Gron ingen, Netherlands: Boekwerk.

Hieronymus, A. N., & Hoover, H. D. (1987). *Iowa Tests of Basic Skills: Writing supplement teacher's guide.* Chicago: Riverside.

Kane, M. T. (1987). Is predictive validity the gold standard or is it the Holy Grail of examinations in the professions? *Professions Education Researcher Notes, 9,* 9-13.

Klass, D., Campbell, C, Hassard, T., Kopelow, M., & Schnabl, G. (1990). Influence of level of training on performance in a standardized test of clinical abilities. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, & R. P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 327-332). Groningen: Netherlands: Boekwerk.

Klass, D., Hassard, T., Kopelow, M., Tamblyn, R., Barrows, H., & Williams, R. (1987). Portability of a multiple station, performance based assessement of clinical competence. In I. R. Hart & R. M. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 434-442). Montreal: Can-Heal.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher,* 20(8), 15-24.

MacRae, H. M. (1991). *The use of checklists as a measure of clinical competence: A validation study.* Unpublished master's thesis. Southern Illinois University School of Medicine, Springfield.

McGuire, C. H. (1983). Evaluation of student and practitioner com-

petence. In C. H. McGuire, R. O. Foley, A. Gorr, R. W. Richardson, & Associates (Eds.), *Handbook of health professions education.* San Francisco: Jossey-Bass.

Mehrens, W. A. (1992). Using performance assessments for accountability purposes. *Educational Measurement: Issues and Practice,* 11(1) 3-9.

Morgan, M. K., & Irby, D. M. (1978). *Evaluating clinical competence in the health professions.* St. Louis: C. V. Mosby.

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

Neufeld, V. R., Woodward, C. A., & Norman, G. R. (1983). Simulated patients in evaluation of medical education. *Proceedings of the 22nd Conference on Research in Medical Education,* 240-242.

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

Newble, D. I., & Swanson, D. B. (1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education,* 22, 335-341.

Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education, 19,* 43-47.

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

Norman, G. R., Feightner, J. W., Tugwell, P., Muzzin, L. J., & Guyatt, G. (1983). The generalizability of measures of clinical problem-solving. In *Proceedings of the 22nd Conference on Research in Medical Education* (pp. 110-114). Washington, DC: Association of American Medical Colleges.

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

Norman, G. R., & Tugwell, O. (1982). A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 53,* 55-58.

Owen, A., & Winkler, R. (1974). General practitioners and psychosocial problems: An evaluation using pseudopatients. *Medical Journal of Australia, 2,* 393-398.

Page, G., Bordage, G., Harasym, P., Bowmer, I., & Swanson, D. (1990). A revision of the Medical Council of Canada's qualifying examination: Prior test results. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, & R. P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 403-407). Groningen, Netherlands: Boekwerk.

Page, G. G., & Fielding, D. W. (1980). Performance on PMP's and performance in practice: Are they related? *Journal of Medical Education, 55,* 529-537.

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

Rethans, J. J., Drop, R., Sturmans, F., & van der Vleuten, C. (1991). A method for introducing standardized (simulated) patients into general practice consultations. *British Journal of General Practice, 41,* 94-96.

Rethans, J. J., Sturmans, F., Drop, R., & van der Vleuten, C. (1991). Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *British Journal of General Practice, 41,* 97-99.

Rethans, J., & van Boven, C. (1987). Simulated patients in general practice: A different look at the consultation. *British Medical Journal, 294,* 809-812.

Reznick, R., Smee, S., Rothman, A., Chalmers, A., Swanson, D., Dufresne, L., Lacombe, G., Baumber,]., Poldre, P., Levasseur, L., Cohen, R., Mendez, J., Patey, P., Boudreau, D., & Berard, M. (1992). An objective structured clinical examination for the licentiate: Report of the pilot project of the Medical Council of Canada. *Academic Medicine, 67,* 487-494.

Rothman, A. I., Cohen, R., Dirks, F. R., & Ross, J. (1990) Evaluating the clinical skills of foreign medical school graduates participating in an internship preparation program. *Academic Medicine, 65,* 391-395.

Rutala, P. J., Witzke, D. B., Leko, E. O., & Fulginiti, J. V. (1991). The influences of student and standardized patient genders on scoring in an objective structured clinical examination. *Academic Medicine, 66,* S28-S30.

Rutala, P. J., Witzke, D. B., Leko, E. O., Fulginiti, J. V., & Taylor, P. J. (1991). Sharing of information by students in an objective structured clinical examination. *Archives of Internal Medicine, 151,* 541-544.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher,* 21(4), 22-27.

Shavelson, R. J., Mayberry, P., Li, W., & Webb, N. (1990). General izability of job performance measurements: Marine Corps infantryman. *Military Psychology, 2,* 129-144.

Shirar, E., Vu, N. V., Colliver, J. A., & Barrows, H. S. (1992). A survey of study methods, preparation time, test-taking strategies, and perceptions of test validity on a clinical performance-based examination. *Academic Medicine, 67,* S10-S12.

Stiggins, R. J., & Bridgeford, N. J. (1986). Student evaluation. In R. A. Berk (Ed.), *Performance assessment: Methods and applications.* Baltimore: John Hopkins University Press.

Stillman, P. L., & Gillers, M. A. (1986). Clinical performance evaluation in medicine and law. In R. A. Berk, *Performance assessment: Methods and applications* (pp. 393-445). Baltimore: John Hopkins University Press.

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

Stillman, P., Haley, H. A., Sutnick, A. I., Philbin, M. M., Smith, S. R., O'Donnell, J., & Pohl, H. (1992). Is test security an issue in a multistation clinical assessment?—A preliminary study. *Academic Medicine, 66,* S25-S27.

Stillman, P., Regan, M. B., & Swanson, D. A. (1987). Diagnostic fourth-year performance assessment. *Archives of Internal Medicine, 19,* 1981-1985.

Stillman, P. L., Regan, M. B., Swanson, D. B., & Haley, H. A. (1992). Gender difference in clinical skills as measured by an examination using standardized patients. In I. R. Hart, R. M. Harden, & J. E. DesMarchais (Eds.), *Current developments in assessing clinical competence.* Montreal: Can-Heal.

Stillman, P. L., Swanson, D. B., Smee, S., Stillman, A. E., Ebert, T. H., Emmel, V. S., Caslowitz, J., Greene, H. L., Hamolsky, M., Hatem, C, Levenson, D. J., Levin, R., Levinson, G., Ley, B., Morgan, G. J., Parrino, T., Robinson, S., & Willms, J. (1986a). Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine, 105,* 762- 771.

Stillman, P. L., Swanson, D. B., Smee, S., Stillman, A. E., Ebert, T. H., Emmel, V. S., Caslowitz, J., Greene, H. L., Hamolsky, M., Hatem, C, Levenson, D. J., Levin, R., Levinson, G., Ley, B., Morgan, G. J., Parrino, T., Robinson, S., & Willms, J. (1986b). *Psychometric characteristics of standardized patients for assessment of clinical skills* (final report to the American Board of Internal Medicine). Philadelphia: American Board of Internal Medicine.

Sutnick, A. I. (1991). *Clinical competence assessment study* (interim report). Philadephia: Educational Commission for Foreign Medical Graduates.

Swanson, D., & Norcini, J. (1989). Factors influencing the reproducibility of tests using standardized patients. *Teaching and Learning in Medicine, 1,* 158-166.

Swanson, D. B., Norcini, J. J., & Grosso, L. T. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12,* 220-246.

Sweeney, J., & Manatt, R. P. (1986). Teacher evaluation. In R. A. Berk (Ed.), *Performance assessment: Methods and applications.* Baltimore: John Hopkins University Press.

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

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

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients. *Teaching and Learning in Medicine, 2,* 58-76.

van der Vleuten, C, van Luijk, S., Ballegooijen, A., & Swanson, D. (1989). Training and experience of medical examiners. *Medical Education, 23,* 290-296.

van der Vleuten, C, van Luijk, S., & Swanson, D. (1988). Reliability (generalizability) of the Maastrict Skills Test. *Proceedings of the 27th Annual Research in Medical Education Conference,* 228-233.

Verhulst, S. J., Colliver, J. A., Paiva, R. E. A., & Williams, R. G. (1986). A factor analytic study of performance of first-year residents. *Journal of Medical Education, 61,* 132-134.

Vu, N. V. (1979). Medical problem solving assessment: A review of

methods and instruments. *Evaluation and the Health Professions, 2,* 282-307.
Vu, N. V., Barrows, H. S., Marcy, M. L., Verhulst, S. J., Colliver, J. C, & Travis, T. A. (1992). Six years of comprehensive, clinical performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Academic Medicine, 67,* 42-50.
Vu, N. V., Colliver, J. A., & Verhulst, S. J. (1992). Factor Structure of Clinical Competence as Assessed in a Performance-Based Examination Using Standardized Patients. In I. R. Hart, R. M. Harden, & J. E. DesMarchais (Eds.), *Current Developments in Assessing Clinical Competence* (pp. 411-415). Montreal: CanHeal.
Vu, N. V,, Distlehorst, L., Verhulst, S. J., & Colliver, J. A. (1993). Performance-based test sensitivity and specificity in predicting first-year residency performance. *Academic Medicine, 68,* S41-S46.
Vu, N. V., Lee, P. A., & Steward, D. E. (1990). Variant accuracy in methods of assessing clinical performance. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, & R. P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 190-194). Groningen, Netherlands: Boekwerk.
Vu, N. V., Marcy, M. L., Colliver, J. A., Verhulst, S. J., Travis, T. A., & Barrows, H. S. (1992). Checklist characteristics and length of testing: Their effects on standardized patients' simulations. *Journal of Medical Education, 26,* 390-395.
Vu, N. V., Marcy, M. L., Verhulst, S. J., & Barrows, H. S. (1990). Generalizability of standardized patients' satisfaction ratings of clinical encounter with fourth-year medical students. *Academic Medicine, 65,* S29-S30.
Vu, N. V., Steward, D., & Marcy, M. (1987). Assessment of the consistency and accuracy of standardized patients' simulations. *Journal of Medical Education, 62,* 1000-1002.
Webster, G. D. (1989). Executive summary on the patient satisfaction questionnaire project. Philadelphia: American Board of Internal Medicine.
Webster, G. D., Shea, J. A., Norcini, J. J., Grosso, L. J., & Swanson, D. B. (1988). Strategies in comparison of methods for scoring patient management problems. *Evaluation and the Health Professions, 11,* 231-248.
Williams, R., Barrows, H., Vu, N., Verhulst, S., Colliver, J., Marcy, M., & Steward, D. (1987). Direct, standardized assessment of clinical competence. *Medical Education, 21,* 482-489.
Williams, R., Lloyd, J. S., & Simonton, D. K. (1992). Sources of OSCE examination Information and perceived helpfulness: A study of the grapevine. In I. R. Hart, R. M. Harden, & J. E. DesMarchais (Eds.), *Current developments in assessing clinical competence.* Montreal: CanHeal.

Received October 22, 1992
Final revision received June 18, 1993
Accepted June 29, 1993

Call for Comments on the **AERA/APA/NCME** *Standards for Educational and Psychological Testing*

The Standards for Educational and Psychological Testing have become the definitive guidelines for the development and use of educational and psychological tests. The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have collaborated for almost 30 years in producing these standards, which are widely accepted by professionals from various disciplines. Separate standards were first issued by APA in 1954, and by AERA and NCME in 1955. The three associations collaborated on subsequent revisions, released in 1966, 1974, and 1985.

Recognizing many recent advances in testing and evaluation, AERA, APA, and NCME recently established a Joint Committee to undertake a revision of the *Standards.* The presidents of the three sponsoring organizations have appointed Eva Baker and Charles D. Spielberger as co-chairs of the Joint Committee that will include 14 additional committee members appointed via consensus of the presidents of the three sponsoring organizations. The Joint Committee welcomes your input in regard to the overall scope and content of the current *Standards,* as well as specific suggestions for modifications or additions.

To facilitate the Joint Committee's review of recommendations, comments must (a) refer to specific existing standards from the 1985 edition [except in those cases where new standards or sections have been proposed]; (b) propose specific wording for new standards or modifications to existing standards; and (c) include a rationale for each proposed modification or addition. Please forward your comments to: Testing Standards Revision Project, Science Directorate, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242, Internet: SCI.APA@EMA1L.APA.ORG, 202-336-6000, FAX202-336-5953. Comments must be received by July 1, **1994,** for consideration by the Joint Committee. Copies of the 1985 *Standards* can be purchased from the APA Order Department at 202-336-5510.

1993 AERA Educational Videotapes
Publishing Qualitative Research
(4 hrs.,set of 2 VHS tapes)

Rodman Webb, University of Florida, leads a 4-hour minicourse during the Annual Meeting. Featured presenters: Lyn Corno, Columbia University; Catherine Emihovich, Florida State University; Richard Wisniewski, University of Tennessee; Elizabeth Bondy, University of Florida; Paul Atkinson, University of Wales; and Mitch Allen, Sage Publications.

- Designed for graduate students and beginning faculty who want to learn more about writing up qualitative research findings and getting their work published in professional journals.

- Features a panel that includes present and past editors of major education journals, experienced reviewers, a book editor, and skilled qualitative researchers.

- Addresses topics such as what editors and reviewers look for in a manuscript, how editors select articles for publication, publishing and tenure, how you can make your writing more effective, and what you should consider when revising a manuscript.

Ordering Information: Prices are $49.00 for AERA members, $78.00 for nonmembers, and $98.00 for institutions. Add $3.00 per item for shipping and handling. DC residents add 6% sales tax. If order is not prepaid, a purchase order must be provided (and an invoicing fee of $4.00 will be added). Order from: **AERA Videotape Sales, 1230 17th Street, NW, Washington, DC 20036-3078.**