

---

# STATUS OF STANDARDIZED PATIENT ASSESSMENT

---

## Methodological Issues in the Use of Standardized Patients for Assessment

**John Norcini**

*Foundation for the Advancement of International Medical Education and Research (FAIMER)  
Philadelphia, Pennsylvania USA*

**John Boulet**

*Research & Evaluation  
Educational Commission for Foreign Medical Graduates  
Philadelphia, Pennsylvania USA*

*Teaching and Learning in Medicine*, 15(4), 293–297

Copyright © 2003 by Lawrence Erlbaum Associates, Inc.

Standardized patient (SP) methodology has changed the way medical students are educated and evaluated. Since Barrows and Abrahamson<sup>1</sup> proposed their use, SPs have achieved broad acceptance and have been the focus of an enormous amount of research. They are now a routine part of the assessment of the clinical skills of medical students, and they have become a key component in several high-stakes certification and licensure programs.<sup>2–4</sup>

In this article, we have selected four issues that we believe are relevant to the status of SPs used in assessment: (a) sources of measurement error, (b) scoring, (c) equivalence, and (d) standard setting. For each, we briefly describe the issue and then offer our views on the state of the research in the area.

### Sources of Measurement Error

Like many other performance-based assessments, the scores or ratings from SP examinations can be subject to a wide array of potential, often interacting, sources of measurement error. Although some of these sources of variability in scores can be minimized through effective examination design, they cannot be eliminated totally. As a result, depending on the amount of testing time available and the structure of the individual exercises, scores from most SP examinations are less reliable than those based on other item formats.

Numerous studies have explored task, scorer, and patient sources of score variability in SP-based assessments.<sup>5–8</sup> For the most part, these investigations have concluded that task variability is the major limiting factor on the reliability of the SP scores. Because

examinees perform differentially depending on the patient, context, and medical content of the encounter, scores on one case are not highly predictive of scores on another. For example, an examinee may be able to take a very good patient history in an area where he or she knows the content area and be less comfortable, and proficient, in completing this task in a situation where she or he is less familiar with the medical condition being simulated. Unfortunately, aside from significantly increasing the number of cases in an examination, there are few other viable solutions to this problem.

Often, test developers attempt to make the cases very generic, focusing simply on the skills to be measured (e.g., history taking, physical examination) rather than the medical content. Although this strategy may slightly enhance reliability, it can also compromise the validity of the assessment by limiting or distorting the medical content that is included. Overall, if score reliability is a major concern, as would be the case for high-stakes SP examinations, it is imperative that examinees be given the opportunity to display their skills across a sufficient number and breadth of encounters.

Although task sampling variability can have a major impact on the reliability of SP assessment scores, it is not the only potential source of measurement error. For example, if examinee performances are evaluated by different groups of scorers (e.g., physicians) then some variability is likely attributable to them, as opposed to true differences in ability between examinees. However, based on the prevailing performance assessment literature,<sup>9</sup> rater effects tend to be small when compared to those associated with the tasks or cases. As a result, provided that rater training is adequate, it is un-

---

Correspondence may be sent to John J. Norcini, Ph.D., President, Foundation for the Advancement of International, Medical Education and Research (FAIMER), 3624 Market Street, 4th Floor, Philadelphia, PA 19104, USA. E-mail: jnorcini@ecfmg.org

likely that appreciable gains in reliability can be achieved by increasing the number of ratings per case.

One important, and often overlooked, component of SP examinations is the choice of patient that plays the role. Often, it is necessary to train more than one SP for a case, and it is assumed that with proper selection and training of these individuals, the cases can be treated as equivalent. Unfortunately, regardless of recruiting strategies, standardized protocols, and SP scripts, this premise may not hold for certain types of encounters. Even minor differences between SPs in terms of body type, pre-existing physical conditions, hygiene, affect, and portrayal may lead examinees to pursue different lines of questioning. As a result, a case may appear to be more difficult when portrayed by one SP as opposed to another. The magnitude of this effect is unlikely to be as large as the variability attributable to completely different cases, but if it is not accounted for in the test design and scoring model, then the accuracy of any competency-based decisions may be adversely affected.<sup>9,10</sup>

### Scoring

One of the enduring arguments about the scoring of SP examinations involves the intertwined issues of the scorers and the forms they complete.<sup>11,12</sup> Some groups use physician scorers, whereas others use the SPs to capture examinees' performances. Similarly, many groups ask their scorers to complete checklists, whereas others prefer global ratings. The choice of scoring model depends to some extent on the purpose of the assessment, the nature of the examinee group, and the logistics of test administration.

The task of the scorer when completing a checklist is to indicate whether a particular behavior occurred. For example, in a case of back pain the scorer indicates whether the examinee asked about the onset of pain, its duration, and the like. Checklist responses will often be dichotomous (1–0 or yes–no), and they serve as an “answer key” for the case, specifying the important and indicated behaviors.

The task of the scorer when completing a rating form is to make a global judgment about the quality of a performance. Again in the back pain case, the scorer might be asked to assess the appropriateness of the history or the soundness of the clinical reasoning. Global ratings are usually collected on a multipoint scale, and scorers must make distinctions among various levels of performance. If the raters have sufficient clinical expertise, this task is relatively straightforward, and valid assessments can be obtained.

There is a sizeable body of research on the use of checklists and rating scales in SP methodology and in the broader literature. Not surprising, scores based on checklists are highly correlated with scores based on

global rating scales suggesting that they are assessing predominately the same aspects of competence. Although the use of checklists may be perceived to be more objective, and can produce slightly more reliable scores, increasing levels of medical expertise may not be captured.<sup>13,14</sup> In general, scores based on global ratings that have been completed by experts tend to be slightly more valid as determined by correlations with other markers of competence.<sup>14–17</sup> However, the overall differences are relatively small so either way of gathering data, provided that the scorers are well trained, produces reasonably valid results.

Decisions about whether to use physicians or SPs as scorers are often based on considerations other than psychometrics. Some groups prefer and need the credibility added by using physicians as scorers. For others, this is impractical and quite expensive. From a psychometric perspective, either physicians or patients can effectively use checklists because they require only that the behavior be noted. In contrast, for global ratings the scorer must be an expert in the competence being assessed to discern differences in quality. It is not unreasonable to view physicians as experts in the medical aspects of the encounter and the patients as experts in the interpersonal and communication aspects.<sup>16,17</sup>

### Equivalence

Different versions or forms of the same SP examination are needed for a variety of reasons, including administration over time and sites. If the versions are not identical in difficulty or the ability to discriminate along the competence continuum, their scores are not equivalent nor are the decisions that emanate from them (e.g., pass–fail decisions). To ensure equivalence, the content of the different tests should conform to the same table of specifications; their difficulty, discrimination, and reliability should be comparable; and some form of statistical adjustment or equating should be applied.<sup>18</sup> SP examinations pose several challenges in this regard.<sup>8</sup>

The simplest and most commonly employed way of ensuring equivalence is to construct different forms of the examination according to the same table of specifications. This is not an acceptable approach for high-stakes tests, especially when the focus is a broad domain such as internal medicine. For example, cardiology cases and even cases of chest pain differ considerably in difficulty depending on a host of characteristics, including patient history, acuity, and the presence of additional symptoms. Therefore, without some form of score adjustment it is not possible to ensure that they will provide equally valid scores. This is a problem for all performance tests, but it is especially acute for SP examinations where the number of cases on the test is usually relatively small.

A number of ways to adjust scores have been developed and used successfully with examinations composed of multiple-choice questions. The methods fall into three general categories: equipercentile, linear, and Item Response Theory (IRT)-based methods.<sup>19,20</sup> They make different assumptions about the forms to be equated so they are suitable for different situations. If the assumptions are not met, however, application of the methods may actually do more harm than good.

In the context of SPs, traditional methods for adjusting scores face significant hurdles. To produce precise results, such methods typically require large numbers of examinees and items (i.e., patients), both of which are unusual in SP examinations. Further, the assumptions of some of the methods may be more difficult to meet. For example, the Rasch model assumes the test is unidimensional and that items and cases are of equal discrimination. It is sometimes difficult to meet these assumptions with a multiple-choice question examination, so their application to a performance test may be even more problematic. Additional research in this area, especially with reference to the dimensionality of SP examination scores,<sup>21</sup> is certainly warranted.

To date, equating has been used only rarely with SP examinations, and the applications have been limited to some of the simpler methods.<sup>8</sup> Considerable research is needed in three areas. First, the effects of examinees and cases should be studied with special attention to how the characteristics of these two aspects of SP examinations interact with the different methods of score adjustment. Second, the fit between some of the scoring models and the data needs to be investigated in the hope that such work will lead to broad guidelines that inform practice. Third, new methods of establishing equivalence should be studied so as to overcome some of the limitations of the traditional approaches.

### Standard Setting

With the increasing use of performance-based assessments as part of professional certification and licensure, considerable effort has been aimed at developing and validating standard setting methods that can be used to make reliable decisions regarding the proficiencies of examinees.<sup>22-24</sup> For SP examinations given in the context of individual medical schools, the purpose of the assessment is often formative rather than summative so there is little need to determine a point on the score scale that can be used to classify examinees into defined proficiency cohorts. Where summative assessment is required in this lower stakes setting, relative standards (e.g., passing the top 90% of examinees) are often better because test quality varies over time and these methods remove differences in difficulty.

In a high-stakes, SP-based examination absolute standards are preferred because relative standards may result in passing examinees without regard to how much they actually know or can do. This is certainly at odds with the purpose of a test of competence. Moreover, relative standards will vary over time with the ability of the examinees creating “vintages” of licensed or certified physicians.<sup>25</sup>

There are a variety of absolute standard-setting procedures that can and have been used for SP assessments.<sup>26-31</sup> These methods can be broadly classified as either test centered or examinee centered. In general, the test-centered methods require subject matter experts to make judgments regarding the expected performance of minimally competent examinees on select tasks. The Angoff<sup>32</sup> procedure and associated modifications can be used to set standards on the history taking and physical examination checklists typically used for scoring cases in SP examinations. Here, the panelists would be required to make judgments as to the probability of a minimally qualified examinee asking the particular question or performing (correctly) the indicated physical examination maneuver. These judgments can then be averaged over panelists and checklist items to obtain a standard for the case.

Examinee-centered methods can also be used to establish standards for SP-based assessments. Instead of judging the test materials such as the checklist items, the panelists view a series of examinee performances and make judgments about which demonstrate proficiency. The task involves distinguishing qualified from unqualified examinees, or simply identifying “borderline” performances. For the former, known as the contrasting group method, the intersection of the distribution of qualified and unqualified examinees can be used to delimit the standard. For the latter, the mean of the scores for the borderline group, or some other measure of central tendency, would define the cutpoint.

The choice of standard setting method to be used for an SP examination will, to some extent, depend on the purpose of assessment and the availability of resources to conduct the exercise. Although test-centered approaches have often been used,<sup>33</sup> it is common for experts to be uncomfortable making judgments about the performance of a hypothetical group of borderline examinees. In contrast, as required by the examinee-centered methods, they are much more at ease making judgments about specific performances. They find the process and results more credible and, because the judgments are based on the actual test performances, their decisions are informed by how the examinees actually did on the test. For these reasons, at least for SP examinations with a minimal number of relatively long tasks, the use of examinee-centered standard setting methods is growing.<sup>22</sup>

## Summary

In summary, we identified four issues that we believe are relevant to the status of SPs used for assessment:

- Cases (i.e., tasks), scorers, patients, and their interactions are sources of variability in examinee scores. In general, cases are the largest contributor, meaning that SP examinations need to have a broad sampling of tasks over multiple patient encounters. Scorers and patients contribute less to error, and increasing the number of cases will reduce their influence as well.

- The intertwined issues of the scorers and the forms they complete attract more attention than they deserve. The overall differences between checklists and global ratings are relatively small. Provided the scorers are well trained, either way of gathering data is reasonable. From a psychometric perspective, either physicians or patients can effectively use checklists, but for global ratings the scorer should be an expert in the domain being assessed.

- In the context of high-stakes assessment, it is important that different versions of a test be of equal difficulty. SP-based examinations pose special challenges in this regard, and considerable research is needed to better understand (a) the effects of the characteristics of examinees and cases on the different methods of score adjustment, (b) the fit between scoring models and the data, and (c) the utility of any new methods.

- The exact choice of method for setting standards on SP-based examinations depends on a number of factors. Nonetheless, examinee-centered methods have grown in popularity because experts are more comfortable making the judgments that underlie them, and their decisions are informed by the actual test performance of examinees.

## References

1. Barrows HS, Abrahamson S. The programmed patient: A technique for appraising student performance in clinical neurology. *Journal of Medical Education* 1964;39:802–5.
2. Brailovsky CA, Grand'Maison P, Lescop J. A large-scale multicenter objective structured clinical examination for licensure. *Academic Medicine* 1992;67(10, Suppl.):S37–9.
3. Whelan GP. Educational commission for foreign medical graduates: Clinical skills assessment prototype. *Medical Teacher* 1999;21:156–60.
4. Medical Council of Canada. *Qualifying examination Part II, information pamphlet*. Ottawa, Ontario; Author, 2002.
5. van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991;25:110–8.
6. Swanson DB, Norcini JJ. Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 1989;1:158–66.
7. Boulet JR, Friedman B-DM, Hambleton RK, Burdick WP, Ziv A, Gary NE. An investigation of the sources of measurement error in the post-encounter written scores from standardized patient examinations. *Advances in Health Sciences Education* 1998;3:89–100.
8. Swanson DB, Clauser BE, Case SM. Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education* 1999;4:67–106.
9. Colliver JA, Swartz MH, Robbs RS, Lofquist M, Cohen D, Verhulst SJ. The effect of using multiple standardized patients on the inter-case reliability of a large-scale standardized-patient examination administered over an extended testing period. *Academic Medicine* 1998;73(10, Suppl.):S81–3.
10. Swartz MH, Colliver JA, Robbs RS, Cohen DS. Effect of multiple standardized patients on case and examination means and passing rates. *Academic Medicine* 1999;74(10, Suppl.):S131–4.
11. Brennan RL, Johnson EG. Generalizability of performance assessments. *Educational Measurement: Issues and Practice* 1995;Winter:9–12.
12. Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? *Academic Medicine* 1996;71:170–5.
13. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenge of creating new OSCE measures to capture the characteristics of expertise. *Medical Education* 2002;36:742–8.
14. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine* 1999;74:1129–34.
15. Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. *Advances in Health Sciences Education* 1997;1:215–9.
16. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine* 1998;73:993–7.
17. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: Comparing checklist and global rating scores. *Academic Medicine* 1999;74(10, Suppl.):S135–7.
18. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 1999.
19. Petersen NS, Kolen MJ, Hoover HD. Scaling, norming, and equating. In RL Linn (Ed.), *Educational measurement* (pp. 221–62). Phoenix, AZ: Oryx, 1993.
20. Kolen MJ, Brennan RL. *Test equating: Methods and practices*. New York: Springer-Verlag, 1995.
21. De Champlain AF, Klass DJ. Assessing the factor structure of a nationally administered standardized patient examination. *Academic Medicine* 1997;72(10, Suppl.):S88–90.
22. Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education* 1999;4:195–207.
23. Meara KC, Hambleton RK, Sireci SG. Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review* 2001;Summer:17–23.
24. Chinn RN, Hertz NR. Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education* 2002;15:1–14.
25. Norcini JJ, Guille R. Combining tests and setting standards. In GR Norman, CPM van der Vleuten, DI Newble (Eds.), *International handbook of research in medical education* (Part 2, pp. 811–34). Dordrecht, the Netherlands: Kluwer Academic, 2002.
26. Norcini JJ, Stillman PL, Sutnick AI, et al. Scoring and standard setting with standardized patients. *Evaluation and the Health Professions* 1993;16:322–32.

27. Dauphinee WD, Blackmore D, Smee S, Rothman AI, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in Health Sciences Education* 1997;2:201–11.
28. Ross LP, Clauser BE, Margolis MJ, Orr NA, Klass DJ. An expert-judgment approach to setting standards for a standardized-patient examination. *Academic Medicine* 1996;71(10, Suppl.):S4–6.
29. Kent A. Setting standards for an objective structured clinical examination: The borderline group method gains ground on Angoff. *Medical Education* 2001;35:1009–10.
30. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: Use of global ratings of borderline performance to determine the passing score. *Medical Education* 2001;35:1043–49.
31. Humphrey-Murto S, MacFadyen JC. Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Academic Medicine* 2002;77:729–32.
32. Angoff WH. Scales, norms and equivalent scores. In RL Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington DC: American Council on Education, 1971.
33. Cusimano, MD. Standard setting in medical education. *Academic Medicine* 1996;71(10, Suppl.):S112–20.

*Received 29 April 2003*

*Final revision 11 October 2003*