The Use of Standardized Patients in Medical Education

Mark H. Swartz, M.D.
Professor of Medicine and
Morchand Professor of Medical Education
Mount Sinai School of Medicine – New York

June 17, 2002
Superior doctors prevent the disease.
Mediocre doctors treat the disease before evident.
Inferior doctors treat the full blown disease.

–Huang Dee: Nai-Ching (2600 B.C. 1st Chinese Medical Text.)
The Use of Standardized Patients in the Teaching and Assessment of Clinical Competence

Mark H. Swartz, M.D.
Mount Sinai School of Medicine – New York
Standardized patients

Background information - 1

- **Standardized patients** (SPs) are being used increasingly worldwide to assess the clinical skills and competencies of medical students, residents, and physicians.

- SP assessments are being used and are under development for licensure and certifying examinations in the United States and Canada.
The thinking underlying this approach is that clinical performance is the ultimate goal of medical education.

Thus, the focus of assessment in medical education has shifted from simply testing knowledge of medical science to testing performance in the clinical setting.
In conjunction with this increased usage of standardized patient testing, a considerable amount of research has been conducted on the psychometric properties of standardized patient examinations.

This research is the primary focus of my talk today.
Standardized patients

History and Definition

- First introduced by Howard Barrows, M.D. and Steven Abrahamson, Ph.D. in 1964
- A standardized patient (SP) is a nonphysician (actor) who has been trained to portray a specific patient case in a consistent, standardized manner.
- In this interaction with the examinee, the SP presents the case history in response to questioning by the examinee and undergoes a physical examination at the examinee’s direction.
- In addition, SPs complete checklists following the examinee-SP encounter, on which they document actions performed by the examinee.
Surveys of SP use

- Survey of 142 curriculum deans of USA and Canada conducted 12 years ago showed that 94 of the 136 deans responding (70%) indicated that SPs were being used in teaching and assessment.

- Many schools reported that they even had a central office that coordinates their SP program.

- Recent study conducted in the past two years showed that nearly all medical schools in the USA and Canada are now using SPs in their curricula.
Licensure and certification

- A number of medical licensing and specialty certifying boards in the USA and Canada are currently using or considering the use of SPs.

- These includes:
  - National Board of Medical Examiners
  - Educational Commission of Foreign Medical Graduates
  - Medical Council of Canada
  - Royal College of Physicians and Surgeons of Canada
  - Corporation of Medical Professionals of Quebec
Psychometric Issues in Standardized Patient Assessment

- **Standardized patient**
  - Realism
  - Accuracy of case portrayal
  - Accuracy of checklist completion

- **Relationships of exam scores with other variables**
  - Gender and ethnicity of SP and examinee
  - Test security
  - Multiple SPs
  - Sequence (or order) effect

- **Reliability**

- **Validity**
Psychometric Issues in Standardized Patient Assessment

- **Standardized patient**
  - Realism
  - Accuracy of case portrayal
  - Accuracy of checklist completion

- Relationships of exam scores with other variables
  - Gender and ethnicity of SP and examinee
  - Test security
  - Multiple SPs
  - Sequence (or order) effect

- Reliability
- Validity
Are standardized patients realistic? Are they accurate in portraying their roles?

- A number of studies have shown that experienced physicians cannot differentiate real patients from SPs when the latter are sent unannounced into a physician’s office. This finding is particularly impressive given that the physicians had agreed in advance to participate in these studies and thus knew that SPs would be coming into their offices unannounced in a period of the next few months.

- Other studies have shown that house officers clinical behaviors were quite similar whether the encounter was with real patients or SPs.
Accuracy of case portrayal

Are they accurate in portraying their roles?

- Several studies have examined SP accuracy
  - On average SPs were more than 90% accurate in portraying the details of the case. Over half had accuracies between 95-100%.
Accuracy of checklist completion

*Are they accurate at checklist completion?*

- As expected, the number of actions needed to be recalled influenced the accuracy.
  - In one study, accuracies ranged from 83% for a 5-item checklist to 76% for a 30-item checklist.
  - At The Morchand Center, accuracies average from 87-98% for a 20-item checklist.
- In general, 15-item checklists appear to represent a good balance between recording accuracy and breadth of information regarding examinee performance.
- Another study showed that SPs reliably evaluated 83% of the same clinical skills that were evaluated by 3 faculty physicians.
Psychometric Issues in Standardized Patient Assessment

- Standardized patient
  - Realism
  - Accuracy of case portrayal
  - Accuracy of checklist completion
- Relationships of exam scores with other variables
  - Gender and ethnicity of SP and examinee
  - Test security
  - Multiple SPs
  - Sequence (or order) effect
- Reliability
- Validity
Gender of SP and examinee

Does the variable of examinee gender with the variable of SP gender facilitate or hinder the performance of examinees?

- Studies have shown no evidence of an interaction of any practical consequence between the variables of examinee gender and SP gender.

- These findings are important because it means that examinees of a given gender will not be favored as a result of the mix of male and female SP cases.
Ethnicity of SP and examinee

Does the variable of examinee ethnicity with the variable of SP ethnicity facilitate or hinder the performance of examinees?

- A related concern is ethnicity; i.e., is there an interaction between examinee ethnicity and SP ethnicity such that examinees would score higher (or lower) depending upon the ethnic background of the SP.

- Research on this question has been difficult to obtain because most SP testing programs lack sufficient numbers of examinees and SPs of different ethnic backgrounds to perform a meaningful statistical analysis.
We were able to address this question with data from the NYC Consortium examination based on two classes of over 1,000 examinees each.

Specifically, we compared white and African-American examinee encounters with that of white and African-American SPs.

Analysis of a number of outcomes for a number of cases showed no significant interactions and no consistent pattern across the different analyses.
Test security

Is test security a potential threat to the validity of SP examinations?

- Studies have shown no consistent, systematic increasing or decreasing trend in scores throughout the examination period.
Test security

- Some SP examinations are typically lengthy.
- At SIU School of Medicine, each class of 70 students is randomly divided into 5 groups of about 14 students each, and the examination in its entirety is administered to one group at a time, requiring 3 days of testing time per group and 15 days for an entire class.

- This schedule raises concerns that students in groups tested earlier might pass on information about the cases to students in groups tested later, in which case, students in the later groups would have an advantage over the students in the earlier groups.

- Fortunately, studies conducted to date have shown no evidence of widespread violations of test security.
Test security

- A study at the University of Illinois at Chicago did show a “grapevine” effect.
- Students do receive information about the examination from other students
- Nevertheless, scores did not improve throughout the examination period.
- Similar informal reports have been obtained from other sites at which test security studies were conducted.
  - Students do talk about the examination but with no systematic improvement in scores throughout the exam period!
Test security

- A study conducted at Mount Sinai went even so far to arrange for students who had been tested on one day to meet with students who would be tested on the next day and to communicate any information about the SP cases that they thought would be helpful.

- Despite this deliberate, somewhat massive assault on test security, performance of students in the second group was unaffected!

- It was concluded that SP examinations by their very nature may be impervious to such violations of test security.
Multiple SPs

Is the reliability of examinees’ total scores affected by the use of multiple SPs on individual cases?

- Because of the time required to administer an SP examination such as to a class of several hundred medical students, two or more (multiple) SPs are commonly used to portray a given case.

- Naturally, this raises the question about the consistency of the simulation from one SP to another.

- Fortunately, studies have shown little or no effect of multiple SPs on examination reliability.
Sequence (or order) effect

*Does the examinee performance improve over a sequence of SP cases?*

- The results are mixed. Three studies have shown an improvement in scores.

- Other studies have not shown any sequence effect.

- Fortunately, the sequence effect, if present, is not a problem for examinees’ total examination scores.
Psychometric Issues in Standardized Patient Assessment

- Standardized patient
  - Realism
  - Accuracy of case portrayal
  - Accuracy of checklist completion
- Relationships of exam scores with other variables
  - Gender and ethnicity of SP and examinee
  - Test security
  - Multiple SPs
  - Sequence (or order) effect
- **Reliability**
- Validity
Reliability

Are scores obtained on SP examinations reliable (reproducible)?

- Reliability is probability the most researched and most documented psychometric property of performance-based examinations that use SP cases.
- Reliability refers to whether measurements obtained with a given measurement process are reproducible with repeated applications of the process.
- It refers to the reproducibility, consistency, or stability of measurements obtained across different trials.
Reliability

- The reliability of a standardized patient examination is typically evaluated with a generalizability coefficient, which can be thought of as a Pearson Correlation Coefficient between scores on the examination at hand and scores on a comparable exam which consists of a different sample of SP cases but from the same population of possible cases.

- Reliabilities of .80 and above are typically recommended for educational and psychological tests.
Reliability

How reliable are raters in the completion of checklists? Is more than one rater per case needed to achieve adequate reliability?

- Studies have concluded that if one can afford additional personnel in an SP examination, the funds should be used to increase the number of cases rather than to increase the number of observers per case. Relatively little precision of measurement is gained by increasing the number of observers per case. Much more is gained by increasing the number of cases.
Reliability

Is reliability (reproducibility) affected by having two or more SPs simulate the same case?

- Two or more SPs are required often to portray a given case.
- Several studies have shown that multiple SPs have little or no effect on examination reliability.
Van der Vleuten and Swanson in a 1990 review reported published score reliabilities from 13 sets of data from SP examinations at 9 institutions. The reliabilities ranged from .41 for the 3-hour exam (Univ. of Texas) to .85 for another 3-hour exam (Univ. of Toronto).
Reliability

Are pass-fail decisions based on SP examinations reliable (reproducible)?

- Typically, the reliability of pass-fail decisions based on an SP examination is of much greater concern than the reliability of the examination scores.

- Fortunately, the reproducibility of pass-fail decisions is much better than that of the examination scores.
Reliability

How many SP cases are needed on an SP examination in order for scores or pass-fail decisions to achieve satisfactory levels of reliability?

- A single score and pass-fail decision based on only one SP case do not provide a reliable assessment of an examinee’s skills and competence. To achieve the recommended .80 reliability, a number of SP cases is needed.

- For 6 years on the NYC Consortium exam, the score reliability was .66; the pass-fail reliability was .98. To get an exam reliability of .80, 15 cases would be needed, but only 2 or 3 cases for a pass-fail reliability of .80.
It is important to note that just as an examination’s pass-fail reliability is typically higher than its score reliability, the number of cases needed to achieve the recommended .80 reliability is typically less for pass-fail reliability than for score reliability.

The key implication of all this is that we should focus more on reliability of pass-fail decisions than on the reliability of examination scores.
Psychometric Issues in Standardized Patient Assessment

- Standardized patient
  - Realism
  - Accuracy of case portrayal
  - Accuracy of checklist completion
- Relationships of exam scores with other variables
  - Gender and ethnicity of SP and examinee
  - Test security
  - Multiple SPs
  - Sequence (or order) effect
- Reliability
- Validity
Validity

Are SP examinations valid?

- Validity refers to whether an examination measures what it purports to measure, and the property or characteristic the examination claims to measure is called a construct.

- Empirical research in support of the validity of SP examinations has been difficult because there is no generally recognized measure of clinical competence to be used as a gold standard criterion.
The validity of an SP examination refers to whether the examination really measures the construct of clinical competence, which would be demonstrated empirically, *by evidence showing that variation in SP-test scores is indicative of actual variation in clinical competence.*

Thus, a measure that closely approximates the clinical competence construct is needed for use as a gold-standard criterion in order to provide convincing evidence of validity of SP examinations.
Validity

- One recurring suggestion for the gold-standard criterion has been global ratings by faculty-physicians observers, although there have been concerns about the reliability of faculty ratings.

- However, the use of faculty ratings is completely consistent with our general reliance upon peer review in the professions for maintaining the standards of the professions and is the basis of major court decisions concerning medical student and resident competency.
The Morchand Center and the NYC Consortium have undertaken the first major validity study of SP assessment using faculty ratings as the gold-standard criterion.
Validity

- Five faculty physicians independently observed and rated videotaped performances of 44 fourth-year medical students in the Class of 1995 from one of the NYC Consortium schools.
- The five raters were from five different schools and did not know any of the students from the sixth school.
- They rated each student’s clinical competence and interpersonal and communication skills using a four-point scale:
  - 4=high pass, 3=pass, 2=low pass, 1=fail
Validity

- The raters were instructed to provide global ratings on the basis of their understanding of the meanings of clinical competence and interpersonal and communication skills.

- The five raters were experienced members of the Consortium’s Steering Committee and were very familiar with the philosophy and purpose of the examination as intended by the Committee.
Validity

- The validity analyses showed correlation between full examination scores and overall ratings from .60 to .70.
- The correlation of the history and physical checklist scores with the overall ratings of clinical competence was .62 and the correlation of the checklist plus post-encounter scores (PES) with the overall ratings was .70. The correlation for interpersonal skills was .65.
- Also, the reliabilities of the faculty ratings were quite good:
  - .82 for clinical competence
  - .77 for interpersonal and communication skills

<table>
<thead>
<tr>
<th>Rating</th>
<th>Hx &amp; PE Checklist</th>
<th>Checklist plus PES</th>
<th>IP plus COM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>.62</td>
<td>.70</td>
<td>.65</td>
</tr>
</tbody>
</table>
Validity

Based on these results, we have been able to report case and examination scores in terms of what might be called “faculty-rating equivalents,” namely, “high pass,” “pass,” “low pass,” and “fail,” or just “pass” and “fail” rather than in terms of the case or examination raw-score scale.
Summary

- Standardized-patient assessment is being used increasingly worldwide, and the research base that accompanies this usage is reassuring, showing that SPs are realistic and are accurate at portraying the patient role and completing checklists, that the examinations can achieve acceptable levels of reliability provided they are of sufficient length, and that relationships with other measures tend to be consistent with our thinking about clinical competence.

- Research on the ultimate, bottom-line question of validity is underway and, even more important, this research is using a format that should enhance our understanding of validity and contribute to the development and refinement of SP assessment.
Thanks to

- The staff of The Morchand Center
  - Devra S. Cohen, *Educational Program Director*
  - Ingrid Baher, *Educational Coordinator*
  - Carol Patterson, *Administrative Assistant*
  - Nidia Williams, *Administrator*

- Faculty members of the NYC Consortium

- Jerry A. Colliver, Ph.D., *SIU School of Medicine*
The Use of Standardized Patients in the Teaching and Assessment of Clinical Competence

Mark H. Swartz, M.D.

Mount Sinai School of Medicine – New York City

mark.swartz@mssm.edu
Thank you for your gracious invitation -

The American Bureau for Medical Advancement in China, Inc.
Types of validity

- **Content validity**
  - Looks at rationale of test construction to determine if the sample of SP cases in the exam is a representative sample of the universe of cases examinees should be able to handle and the competencies that should have been mastered
Types of validity

- **Criterion validity**
  - Looks at correlation of test scores and scores on a gold-standard criterion, like faculty-physician global ratings
Types of validity

- **Construct validity**
  - Given meaning and implications of construct purportedly measured by exam; research is conducted to see if exam scores reflect that meaning and interpretation, e.g., scores increase with training.
  - Or, studies that look at relationships of other variables with exam scores can be said to determine empirically the meaning of the test scores.