

# Correlation between Housestaff Performance on the United States Medical Licensing Examination and Standardized Patient Encounters

WILLIAM D. RIFKIN, M.D.<sup>1</sup>, AND ARTHUR RIFKIN, M.D.<sup>2</sup>

## Abstract

**Background:** There is interest in the use of “standardized patients” to assist in evaluating medical trainees’ clinical skills, which may be difficult to evaluate with written exams alone. Previous studies of the validity of observed structured clinical exams have found low correlation with various written exams as well as with faculty evaluations. Since the United States Medical Licensing Examination (USMLE) results are often used by training programs in the selection of applicants, we assessed the correlation between performance on an observed structured clinical exam and the USMLE, steps 1 and 2, for internal medicine housestaff.

**Methods:** We collected scores on the USMLE, steps 1 and 2, and the overall score from a required standardized patient encounter for all PGY-1 trainees, in a single urban teaching hospital. Pearson coefficients were used to compare the USMLE and observed structured clinical exam performance.

**Results:** The two steps of the USMLE correlated with each other to a large extent ( $r=0.65$ ,  $df=30$ ,  $p=0.0001$ ). However, both steps of the USMLE correlated poorly with the observed structured clinical exam (step 1  $r=0.2$ ,  $df=32$ ,  $p=0.27$ ; step 2  $r=0.09$ ,  $df=30$ ,  $p=0.61$ ).

**Conclusions:** The low correlation between the USMLE and performance on a structured clinical exam suggests that either the written exam is a poor predictor of actual clinical performance, the small window of clinical skills measured by the structured clinical exam is inadequate, or the two methods evaluate different skill sets entirely. Our findings are consistent with previous work finding low correlations between structured clinical exams and accepted common means of evaluation, such as faculty evaluations, other written exams and program director assessments. The medical education community needs to develop an objective, valid method of measuring important, yet subjective, skill-sets such as interpersonal communication, empathy and efficient data collection.

**Key Words:** Standardized patients, observed structured clinical examination, housestaff, United States Medical Licensing Examination (USMLE), performance.

## Introduction

THERE IS INCREASING INTEREST in using “standardized patients” to assist in evaluating the clinical skills of residents, i.e., to reliably and validly measure history taking, physical examinations and interpersonal skills, which are difficult to measure on a written exam (1). More than a decade ago, the American Board of Internal Medicine (ABIM) discontinued the oral portion of the certifying exam and left assessment of actual clinical performance to the individual residency program directors. They, in turn, have employed several methods to fulfill this requirement,

ranging from summation of individual faculty evaluations to formal observation of clinical exercises with patients. It is hoped that by offering a standardized format, there will be a reduction in the bias of non-standard faculty evaluations of differing clinical presentations (2–4).

This approach to evaluating clinical skills has now been used for several years to assess the clinical skills of international medical graduates (IMG) as a requirement for Educational Commission for Foreign Medical Graduates (ECFMG) certification, which is a prerequisite to medical training in the United States. Due in large measure to public support for such objective validations of clinical skills, a Clinical Skills Examination (CSE) will become a part of the licensing requirements for all US medical school graduates. Beginning with the class graduating in 2005, step two of the United States Medical Licensing Examination (USMLE) will include a written portion and an observed standardized clinical examination (OSCE) portion (5).

The perceived necessity of such an external measurement of clinical skills has grown out of concern that medical school faculty have not been assuring such skills, due to lack of direct observation of student performance and faculty members’ varying abilities to detect deficiencies (5).

From the <sup>1</sup>Department of Medicine, Yale University School of Medicine, New Haven, CT, and <sup>2</sup>Department of Psychiatry, Albert Einstein School of Medicine, Bronx, NY and the Zucker Hillside Hospital, Glen Oaks, NY.

The work was done at the Department of Medicine, Maimonides Medical Center, Brooklyn, NY.

Address all correspondence to William Rifkin, M.D., Associate Director, Yale Primary Care Residency Program, Waterbury Hospital, 64 Robbins Street, Waterbury, CT 06721; email: wrifkin@wtbyhosp.chime.org

Grant support: none. Previously presented as a poster at the Society of General Internal Medicine national meeting, April 30 – May 3, 2003, Vancouver, British Columbia.

Accepted for publication July 2004.

Similar concerns have been raised at the Graduate Medical Education (GME) level as well. Specifically, suggestions to date have centered on closer day-to-day observation of patient care and incorporation of the ABIM's Clinical Evaluation Exercises (CEX and mini-CEX) (5–7).

Previous work has approximated the inter-rater reliability of so-called "standardized patients" (4, 8–9), with results ranging from 0.40 to 0.80. Other studies have examined the validity of performance on an observed structured clinical exercise using standardized patients by correlating performance with various so-called "gold standards." Correlations with overall program director assessment of intern performance, as well as summation of faculty evaluations ranged from 0.00 to 0.40 (2, 3). Written exams (in-training exam, ABIM written certifying exam, and its predecessor, the National Board of Medical Examiners Examination) correlated, with a range from 0.22 to 0.30 (2, 4). Correlations were more substantial between written exams, for example, between the in-training exam taken during postgraduate year-2 (PGY-2) and PGY-3 years and the ABIM exam results (0.59 and 0.68, respectively) (10). Previous work found that USMLE step one (0.16) and step two (0.30) correlated modestly with fourth-year medical student performance on an observed structured clinical exam (11).

We assessed whether performance with standardized patients correlates with USMLE scores for PGY 1 internal medicine residents. To our knowledge, this has not been done previously. The result of the study may be of interest especially to program directors, since the USMLE is often used as a major criterion for housestaff selection in the United States (12).

### Methods

We collected scores on the USMLE, steps one (basic science) and two (clinical science), and the overall score from a required standardized patient encounter for all PGY-1 internal medicine housestaff at an urban tertiary-care community teaching hospital, over two years (2001 and 2002). First attempt USMLE results for each part were used in the case of multiple attempts. The standardized patient encounter was administered in the fall of the intern year. Each physician was scored in three domains (history taking, physical examination and interpersonal skills) while assessing four actor-patients portraying the following: obesity, shortness of breath, sickle cell disease and abdominal pain. Stated inter-rater reliability for this exercise was 0.49 in 2002 and 0.62 in 2001 (13). We used Pearson correlation coefficients to compare steps one and two of the USMLE and the standardized patient exam. There was no external funding provided.

### Results

We had 32 pairs of scores available for comparison between the two steps of the USMLE and between step two and the standardized patient exam. We had 34 pairs of scores available for comparison between step one and the standardized patient exam. As expected, the two steps of the USMLE correlate to each other to a large extent ( $r = 0.65$ ,  $df = 30$ ,  $p = 0.0001$ ). But we found that the standardized patient exam correlates poorly with both steps of the USMLE. The relationship to step one ( $r = 0.20$ ,  $df = 32$ ,  $p = 0.27$ ) accounts for only 4% of the variance. The relationship to step two is even less than to step one ( $r = 0.09$ ,  $df = 30$ ,  $p = .061$ ), accounting for only 0.8% of the variance.

### Discussion

The very low correlation found between the United States Medical Licensing Examination and performance on the observed structured clinical exam suggests that either the USMLE is a poor predictor of actual clinical performance, or conversely, that the small window of clinical skills measured by the standardized patient exam affords little ability to discriminate. It was surprising to find that step two of the USMLE, the clinical knowledge section, was not more highly correlated than was step one, the basic science section. And why step one correlated more closely than did step two is not clear. It is possible that this variation was due to chance, that is, the observed differences between step one and two were not significant or meaningful. We could speculate that some aspect of the particular OSCE employed or the population examined, mostly IMGs, could account for this finding. To examine this, we tested the null hypothesis that the step 1 and step 2 correlations to the clinical exam are not different. Our results ( $z=0.4354$ ,  $p=0.33$ ) confirmed this; thus we are unable to conclude that the correlations are indeed different. We therefore conclude that neither step of the USMLE predicts performance on the OSCE.

Some investigators who also found low correlations between standardized patient tests and other forms of evaluation (program director assessments, faculty evaluations, written examinations) suggested that this demonstrates that the different evaluations assess different skills, citing this as an advantage (1, 9, 11).

An assessment demonstrates its usefulness by predicting an important outcome. One of the assessments with which the observed structured clinical exam correlated poorly in this study (basic science written exam scores) does predict obtaining ABIM certification (14). Therefore, does the observed structured clinical exam provide useful predictions? This remains to be demonstrated. To do so requires a "gold

standard” for good clinical skills, and it is unlikely that one such standard will ever appear. Perhaps the optimal procedure involves cross-correlation of several assessments with high face validity, such as real-time faculty or peer assessments.

Limitations of this study include the fact that our sample included many non-US-citizen international medical graduates (45%) and US-citizen international medical graduates (25%), who may not be representative of programs nationwide. In addition, our sample was drawn from a single residency program and may not be replicated in larger, more geographically diverse samples. It should also be noted that while the USMLE steps were taken in medical school, the OSCE was taken during internship, that is, at two distinct phases in medical education. It is possible that some of the variation found could be due to aspects of training incorporated during the latter part of medical school or during the internship itself. However, this would not detract from our conclusion that the USMLE itself is a poor predictor of the OSCE performance and perhaps then more generally of clinical performance, if one presumes that OSCE performance should predict performance in true clinical practice.

In conclusion, brief observations of patient assessments should be tested against a measure of clinical competence that has proven validity. That is, the medical education community needs to develop and test an objective method of measuring important, yet subjective skill-sets such as history-taking, physical examination and interpersonal skills.

#### References

- Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med* 1998; 129(1):42–48.
- Petrusa ER, Blackwell TA, Ainsworth MA. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Arch Intern Med* 1990; 150:573–577.
- Dupras DM, Li JT. Use of an objective structured clinical examination to determine clinical competence. *Acad Med* 1995; 70(1):1029–1034.
- Stillman P, Swanson D, Regan MB, et al. Assessment of clinical skills of residents utilizing standardized patients. A follow-up study and recommendations for application. *Ann Intern Med* 1991; 114(5):393–401.
- Holmboe ES. Faculty and the observation of trainees’ clinical skills: problems and opportunities. *Acad Med* 2004; 79:16–22.
- Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995; 123:795–799.
- Holmboe ES, Fiebach NF, Galaty L, Huot S. The effectiveness of a focused educational intervention on resident evaluations from faculty: a randomized controlled trial. *J Gen Intern Med* 2001; 16:1–6.
- Stillman PL, Swanson D, Smee S, et al. Assessing clinical skills of residents with standardized patients. *Ann Intern Med* 1986; 105:762–771.
- Hull AL, Hodder S, Berger B, et al. Validity of three clinical performance assessments of internal medicine clerks. *Acad Med* 1995; 70:517–522.
- Grossman RS, Fincher RM, Layne RD, et al. Validity of the in-training examination for predicting American Board of Internal Medicine certifying examination scores. *J Gen Intern Med* 1992; 7(1):63–67.
- Swartz MH, Colliver JA, Bardes CL, et al. Validating the standardized-patient assessment administered to medical students in the New York City consortium. *Acad Med* 1997; 72:619–626.
- Wagoner NE, Suriano JR. Program directors’ responses to a survey on variables used to select residents in a time of change. *Acad Med* 1999; 74:51–58.
- Clinical Assessment Examination PGY-1 Housestaff 2002. Aggregate Analyses. Office for Graduate Medical Education, Mount Sinai School of Medicine.
- Sosenko J, Stekel KW, Soto R, Gelbard M. NBME examination part I as a predictor of clinical and ABIM certifying examination performances. *J Gen Intern Med* 1993; 8:86–88.