

## A study of Two Clinical Performance Scores: Assessing the Psychometric Characteristics of a Combined Score Derived from Clinical Evaluation Forms and OSCEs

Clarence D. Kreiter, PhD\*, George R. Bergus, MD, MA(Ed)†

\*Office of Consultation and Research in Medical Education, Department of Family Medicine

†Department of Family Medicine, Performance Based Assessment Program, Office of Student Affairs and Curriculum

Carver College of Medicine, The University of Iowa

Iowa City, Iowa, USA

### **Abstract:**

**Background/Purpose** - It is important to improve the quality of clinical skill assessments. In addition to using the OSCE, the clinical skills of medical students are assessed with clinical evaluation forms (CEFs). The purpose of this study is to examine the psychometric characteristics of an OSCE/CEF composite score.

**Methods** - This study included 2 medical student classes from a large medical school. Students completed approximately 12 OSCEs and were rated over 33 times on a CEF. The reliability of the CEF and OSCE were estimated. A correlation between the mean OSCE score and the mean CEF was calculated and corrected for attenuation. Classical methods were used to examine composite score reliability.

**Results** - For both classes there was a statistically significant correlation between the CEF and OSCE ( $r = .27$  &  $.42$ ,  $p = .003$  &  $.0001$ ). The disattenuated correlations were  $.44$  and  $.68$ . Weighting the OSCE in the composite score as high as  $.4$  was associated with only a small decrease in composite reliability.

**Conclusion** - These results demonstrate that assessment information based on simulated and actual patient encounters can be combined into a composite. Since a composite score may provide a more valid measure of clinical performance, this study supports using a combined CEF and OSCE measure.

**Keywords:** psychometrics, composite score, clinical skills, OSCE, reliability

It is important that students have basic mastery of medical skills upon graduation. To determine if students possess these skills, it is essential that medical schools obtain an accurate measure of each student's clinical competence. In many instances schools rely on OSCE assessments to summarize their students' clinical skill achievement. Although it is difficult to assess the effectiveness nationally of these within-school performance-based assessments, the implementation of clinical skill exams by the USMLE and the Medical Council of Canada suggests a perception that individual schools, at least in North America, cannot certify the skills of their graduates. Since medical schools may not be assessing clinical skills in a way that adequately identifies student deficiencies, it is important for researchers to investigate methods for improving clinical skill assessment systems within the medical college.

Improving the quality of the clinical skill assessments that take place within our medical schools

and implementing remediation for those who fail could assure the public regarding the competency of graduating students. Certainly, medical schools have a strategic advantage over the single clinical skill testing session conducted by the USMLE. Medical schools are capable of more frequently assessing students and doing so in a wider range of contexts with assessments that tap a more multi-faceted array of clinical skills. For example, in addition to using the OSCE, the clinical skills of medical students are also routinely assessed with clinical evaluation forms (CEFs) on which faculty and residents rate students based on direct observation in a clinical setting. It is reasonable to assume that the OSCE and the CEF each measure a somewhat different but related aspect of a student's clinical competence. The OSCE involves observing students in simulated encounters and often provides information about a student's communication skills as well as their skills at collecting clinical data. On the other hand, the CEF is based on the observations of faculty and residents within a real clinical setting and

provides information about a student's clinical reasoning, relevant knowledge, and their ability to interpret clinical data. However, the CEF probably provides limited information about a student's communication skills as interactions with patients are not frequently observed by faculty or resident raters.<sup>1,2</sup> Because the OSCE and the CEF each measure somewhat different aspects of a student's performance, a total clinical skills score that combines the 2 related but different measures may yield a composite score that reflects a more comprehensive assessment of clinical performance. The purpose of this study is to examine and relationship between the OSCE and the CEF. It considers the psychometric characteristics and validity of a composite that combines the OSCE and CEF to generate a final score for summarizing students' clinical skills and making competence-based decisions.

## Methods

This study included 2 cohorts of students from the University of Iowa Carver College of Medicine. In the first cohort there were 122 third year medical students from the class of 2005 who completed 8 or more standardized patient (SP) encounters during the 2003/04 academic year. This group of students was also rated by residents and faculty using a standard CEF an average of 33.54 times (range 18-60) during their clerkship rotations. The second student cohort from the class of 2006 included 134 students who completed 8 or more SP encounters during the 2004/05 academic year. This second cohort was rated an average of 33.69 times (range 14-52) during their clerkship year using the same CEF. All CEF ratings were performed using a standardized 11-item form with each item utilizing a 5-option Likert-type response scale on which 1 represented the lowest level of performance and 5 the highest. OSCEs were distributed across the clerkship year in association with the 5 major clerkships (Surgery, Pediatric, Psychiatry, Obstetrics/Gynecology, Out-Pt. Internal Med. / Family Med.), with some clerkships changing their case offerings during the year. Because not all students took all of the clerkships and because some cases changed during the year, students did not all experience the same group of cases. On average, students experienced 11.01 cases (range 9-12) during the 2003/04 academic year and 12.95 cases (range 8-15) during the second academic year (04/05).

Within each year, OSCE scores were aggregated by first standardizing the scores for each case and then computing an average score across cases for each student. The reliability of this average measure was estimated using an unbalanced generalizability study design. The design and results of the generalizability analysis are reported in detail in an earlier research report.<sup>3</sup> The

reliability of the CEF was estimated using a rater nested within student G study design.<sup>4</sup> A correlation between the mean OSCE score and the mean CEF was calculated, tested for significance, and corrected for attenuation due to the unreliability of both the average CEF score and the average OSCE score by applying the equation below. In the equation,  $r_t$  is defined as the "true score" correlation and is calculated as:

$$r_t = r_{xy} / \sqrt{r_{xx} \times r_{yy}}$$

Where:

$r_t$  = the true score correlation,

$r_{xy}$  = the observed correlation between the mean CEF and the mean OSCE,

$r_{xx}$  = the reliability of the CEF, and

$r_{yy}$  = the reliability of the OSCE.

The equation above allows an estimate of the "true" correlation between the CEF and the OSCE. Or stated another way, it estimates what the correlation would have been had it been computed between perfectly reliable CEF and OSCE measures.

Classical test theory (CTT) methods were used to calculate a composite score reliability using various weightings on the 2 measures. Thissen and Wainer<sup>5</sup> describe the CTT calculation for composite score reliability presented below.

$$r_c = 1 - \frac{\sum_v w_v^2 (1 - r_v)}{\sum_v w_v^2 + \sum_v \sum_{v'} w_v w_{v'} r_{vv'}}$$

Where:

$r_c$  = the reliability of the composite score,

$w_v$  = the weight for component  $v$ ,

$r_v$  = the sample estimate reliability for component  $v$ , and

$r_{vv'}$  = is the correlation between components  $v$  and  $v'$ .

## Results

Table 1 reports the end-of-year average scores, the average number of observations, the standard deviations (SDs), and the estimated reliability for both the CEF and the OSCE for the classes of 2005 and 2006. The mean value of the CEF was very similar across the 2 years. The mean reported values of the OSCE scores are also similar across the 2 years, but this is a result of scaling each case to have a mean of 80 and a SD of 5.0.

Table 2 examines the correlations between the CEF and the OSCE for the 2 classes. In both cohorts there

**Table 1**  
**Descriptive Summary of Performance Based Scores**

Class (Number of Students)	Assessment Type	Average Score	Aver. Obs. Per Stud. (Range)	SD	Reliability
<b>2005</b> (n = 122)	CEF	4.19	33.54 (18-60)	0.18	.70
	OSCE	80.01	11.01 (9-12)	2.25	.54
<b>2006</b> (n = 134)	CEF	4.18	33.69 (14-52)	0.20	.70
	OSCE	80.07	12.94 (8-15)	2.07	.56

was a statistically significant correlation between the 2 mean measures ( $r = .27$  &  $.42$ ,  $p = .003$  &  $.0001$ ). For the class of 2005, the variables displayed somewhat less association compared with 2006, but this difference was not statistically significant ( $z = 1.4$ ,  $p = .16$ ). The correlations corrected for attenuation due to unreliability (the true score correlation) are reported in the last column of Table 2. For 2005 the disattenuated correlation was  $.44$ , while for the class of 2006 it was  $.68$ .

**Conclusion**

These results demonstrate that assessment information based on simulated clinical encounters and actual patient encounters can be combined into a composite measure of performance providing more information than using either singly. The advantages of this synthesis remain theoretical at this time, but the rationale is that these 2 individual scores have their own unique strengths in

**Table 2**  
**Correlation Between the OSCE and CEF for Classes of 2005 and 2006**

Class	Correlation btw CEF & OSCE	Significance	Disattenuated Correlation
<b>2005</b> (n = 122)	$r = .268$	$p = .0028$	$r_t = .44$
<b>2006</b> (n = 134)	$r = .424$	$p = .0001$	$r_t = .68$

Figure 1 displays the composite score reliability for various weighting on the CEF and OSCE when both measures are standardized to a common scale. On the horizontal axis is the weight on the OSCE. The weight on the CEF is equal to 1 minus the OSCE weight. Hence, as the effective OSCE weight goes up, the CEF weight goes down, and vice-versa. For both years, it is clear that the reliability of the composite tends to decrease as the weight on the OSCE increases above  $.25$ . In terms of optimizing the composite reliability, a weight of approximately  $.25$  on the OSCE and  $.75$  on the CEF would be indicated. However, increasing the weighting on the OSCE of up to  $.4$  is associated with only a small decrease in composite reliability.

assessing student clinical competence and that both measured aspects are important in defining clinical competence. The CEF is based on direct observations of a student while working within the process of care delivery and comes primarily from observations of the student while clinical rounding, discussing patients with their supervisors, and working with others to satisfy clinical work demands. Thus, the CEF should provide useful insight into a student's clinical knowledge, ability to work within teams, and professionalism. However, students are infrequently observed by faculty and residents while interacting with patients,<sup>1,2</sup> and these ratings do not inform us about how students communicate with patient and go about collecting clinical information from patients. The OSCE, however, does allow this information to be collected through direct observation of performance in

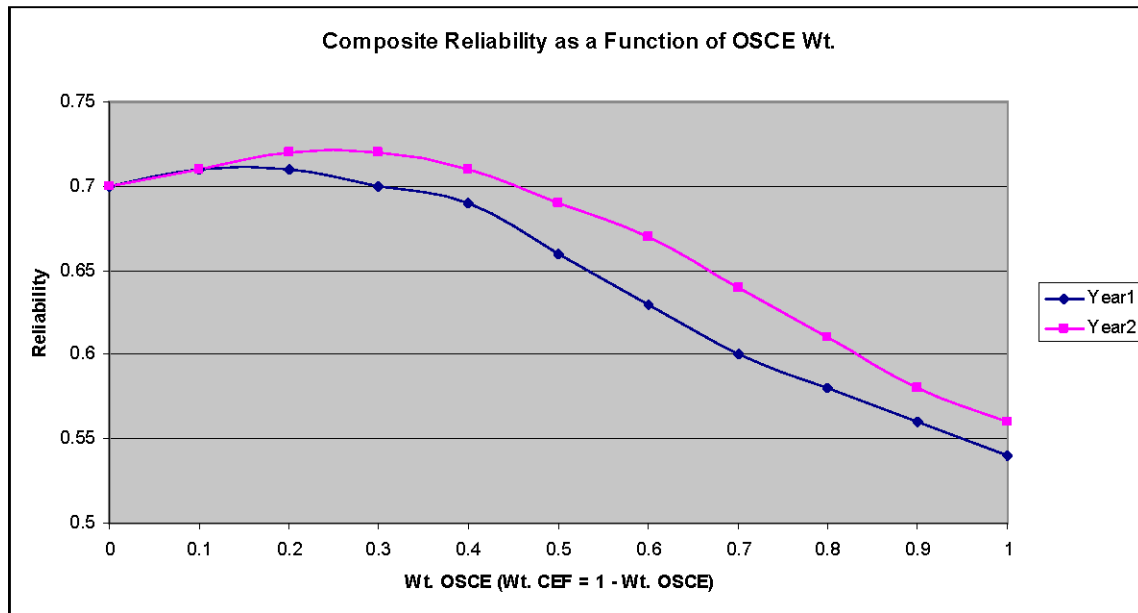


Figure 1 Composite Reliability as a Function of OSCE Weight for Classes of 2005 and 2006

the simulated encounter. An additional contrast between the OSCE and CEF assessment is that OSCE performance data are collected during discrete testing sessions, while the CEF is work-centered. Thus, the OSCE, unlike the CEF, may be incapable of providing information about how a student habitually performs. Miller called attention to this distinction as the difference between the “shows how” and “does” levels of performance in the assessment hierarchy.<sup>6</sup>

We find that, consistent with previous reports in the literature,<sup>7,8</sup> there was moderate correlation between the aggregated CEF score and OSCE score. Unlike these earlier reports, our correlations are based on data collected using multiple clerkships rather than a single clerkship. As the OSCE and CEF focus on different aspects of clinical competency, it is intuitively attractive to study how these partially overlapping measures can be combined to convey information about a more comprehensive definition of clinical competency.<sup>9,10</sup> In addition, by combining the 2 scores it is possible to produce a more reliable score for assigning final grades and making local competency decisions compared with using either measure singly.

A composite score, created by combining CEF and OSCE scores, may provide a more valid measure of clinical performance than either of its components alone. Whether this is the case requires additional study. The magnitude of the correlation between the CEF and OSCE, however, provides important validity evidence. These findings allow an evidence-based response to the

criticism that the CEF may measure only superficial or clinically unimportant dimensions of student performance or that performance on the OSCE has little relevance to practice in clinical settings. This study and others<sup>7,8,11</sup> show the CEF to be positively correlated with important clinically-related measures and strongly support using the CEF for the calculation of clinical grades and for making competency decisions. Such uses for the mean CEF score are further supported by the finding in previous studies<sup>4,12</sup> that the reliability of a mean CEF score when averaged over observations collected across the clerkship year produces an acceptably reliable score.

There are several clear limitations to our study that affect the generalizability of our findings. First, the OSCE scores were obtained from performance assessments embedded within clerkships, while many schools use the OSCE within the context of a single high stakes exam at the end of the clinical year. Whether these scores show a similar correlation with CEF ratings is not known but deserves study. Additionally, not all medical schools use the same CEF on all clerkships. We do not know whether scores from different CEFs can be aggregated to produce a single measure of student performance. Despite these limitations, our study contributes to the literature on clinical evaluation and illustrates the potential arising from combining separate scores into a single measure of clinical competence.

## References

1. Howley LD, Wilson WG. Direct observation of

- students during clerkship rotations: a multi-year descriptive study. *Acad Med.* 2004; 79(3):276-80.
2. Lane JL, Gottlieb RP. Structured clinical observations: a method to teach clinical skills with limited time and financial resources. *Pediatrics*, 2000;105(4 Pt 2):973-7.
  3. Bergus GR, Kreiter CD. The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. *Med Educ.* 2007; 4(7):661-6.
  4. Kreiter CD, Ferguson KJ. Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Eval Health Prof.*2001; 24(1):36-46.
  5. Wainer H, Thissen D. True score theory: the traditional method. In: Thissen D, Wainer H, editors. *Test Scoring*. Mahwah (NJ): Lawrence Erlbaum Associates; 2001. p. 23-72.
  6. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* 1990; 65(9 Suppl):S63-7.
  7. Prislin MD, Fitzpatrick CF, Lie D, Giglio M, Lewis E. Use of an objective structured clinical examination in evaluating student performance. *Fam Med.* 1998; 30(5):338-44.
  8. Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, et al. Validity of three clinical performance assessments of internal medicine clerks. *Acad Med.*1995; 70(6):517-22.
  9. Elstein S. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Acad Med.* 1993; 68(4):244-9.
  10. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001; 357(9260):945-9.
  11. Ferguson KJ, Kreiter CD. Using a longitudinal database to assess the validity of preceptor ratings of clerkship performance. *Adv Health Sci Educ TheoryPract.* 2004; 9(1): 39-46.
  12. Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performance. *Acad Med.* 1998; 73(12):1294-8.

#### Correspondence

Clarence D. Kreiter, PhD  
Office of Consultation and Research in Medical  
Education  
1204 MEB  
The University of Iowa  
Iowa City, IA 52242, USA  
  
Phone: (319) 335-8906  
Fax: (319) 335-8904  
E-mail: [clarence-kreiter@uiowa.edu](mailto:clarence-kreiter@uiowa.edu)