Author's response to reviews

Title: How Well Do Second-Year Students Learn Physical Diagnosis? Observational Study of an Objective Structured Clinical Examination (OSCE)

Authors:

- Claus Hamann (chamann@partners.org)
- Kevin Volkan (kevin.volkan@csuci.edu)
- Mary B Fishman (MBF3@gunet.georgetown.edu)
- Ronald C Silvestri (rsilvest@caregroup.harvard.edu)
- Steven R Simon (steven_simon@hms.harvard.edu)
- Suzanne W Fletcher (suzanne_fletcher@hms.harvard.edu)

Version: 2 Date: 20 Nov 2001

PDF covering letter
November 16, 2001

Carole Mongin-Bulewski, PhD
Editorial Assistant
BioMedCentral
Middlesex House
34-42 Cleveland Street
London W1T 4LB

Tel: +44 (0)20 7323 0323
Fax: +44 (0)20 7636 4296
email: editorial@biomedcentral.com
Web: www.biomedcentral.com

Re: Manuscript:
How Well Do Second-Year Students Learn Physical Diagnosis?
Observational Study of an Objective Structured Clinical Examination (OSCE)

Dear Dr. Mongin-Bulewski:

Thank you for your e-mail of October 30 and the invitation to submit our revised article. We appreciate the reviewers' comments. Our responses to Dr. Mavis follow under A. below, and to Dr. Townsend under B. beginning on p. 6. Manuscript changes are indicated in bold. The full text changes are underlined in the attachment to this letter beginning on p. 6, and the additional references are indicated in the final section of this letter on p.8. We have not changed the tables or figure.

I have electronically submitted our revised manuscript. Thank you for considering our study for publication in BioMedCentral.

Sincerely,

Claus Hamann MD, MS
chamann@partners.org

A. RESPONSE TO REVIEW BY DR. BRIAN MAVIS

Since Dr. Mavis’ had major concerns about inter-rater reliability (his comment # 2) and certain aspects of validity (his comment # 3), we give detailed responses to those comments first, under 1. and 2. below. We respond to his first concern under #3.

1. The lack of inter-rater reliability is a serious shortcoming…nor is there information about how familiar the raters were with the forms being used.

Dr. Mavis points out a legitimate concern about the lack of inter-rater reliability (IRR). This type of reliability gives an indication of whether or not ambiguity in the test items lead different observers to make inconsistent ratings of a student’s performance. In other words, IRR gives an indication of the consistency and repeatability of a test. Some
measurement of these characteristics is needed if the results of a test are to be believed in the context of a research study.

In our study, it was not logistically possible to obtain IRR because of the need for many examiners and because of cost. These logistical problems with OSCE’s are common. It has been suggested that in this situation performance tests using many stations with a single rater may be better than fewer stations with multiple raters (Medical Council of Canada, 1995). It has also been suggested that high IRR developed through the standardized training of a small group of raters will not translate well to large-scale assessments with many raters. As Clauser reports, “Not only is it more difficult to train larger groups of raters to perform in a standardized manner, but issues of recruitment can come into play. It might be relatively easy to recruit five individuals capable of learning to produce accurate ratings; recruiting 50 could be a very different matter.” (Clauser, 2000, p. 316).

To overcome the lack of rater reliability measurement in our OSCE, we used a Generalizability Theory (GT) framework to give measure of reliability similar in quality to IRR (Brennan, 2000). There are a number of examples of the use of GT to statistically account for rater error (Braun, 1988; Clauser, 2000; Englehard, 1994; Longford, 1995). A GT analysis statistically accounts for rater error by parsing out the variance relevant to the instrument in question. By modeling the variances as separate characteristics, we can isolate the variance due to student ability, which in classical test theory is equivalent to true score variance. Other variances related to the test are treated as error variances. In this framework, we can treat error due to differences in raters as error variance.

We described our approach to reliability measurement in the Methods of the original submission, at the top of p.6, paragraph 2. In response to Dr. Mavis’ concerns, we have added referenced text in the Methods, top of p.6 (Attachment, A.1a.), and in the Discussion, beginning at the top of p. 9 and including two new paragraphs (Attachment, A.1b.)

Using GT can overcome some problems inherent in IRR, such as overestimating reliability (Fan & Chen, 2000). Due to the large number of raters involved in the testing, we can assume that any error due to rater differences is randomly distributed. Due to the well-known properties of the central limit theorem in statistics, we know that randomly distributed error has a mean of zero. Therefore, in the GT analysis of our OSCE, we can assume the error variance due to differences among all raters is zero. This assumption is especially appropriate when the test scores used in the analysis are created by summing many items across each scale. Summing in this fashion has the effect of further randomizing the error variance.

Another benefit of the GT approach is that the reliability coefficient derived from the GT analysis is equivalent to Cronbach’s alpha-coefficient which, for binary items, is equivalent to the KR-20 reliability coefficient. The alpha coefficient is especially useful during test development because it gives a measure of how each item is contributing to the scale to which it has been assigned. This measure makes it easy to target items for editing or deletion if they are not performing well. Since we are ultimately interested in
using the scale scores for our research study, this measure of reliability makes sense and is commonly used in OSCE’s.

Checklists have been recommended for use in OSCE over other rating methods such as global ratings, specifically to improve reliability (van der Vleuten & Swanson, 1990). Problems with checklists used in OSCE are usually due to inappropriate design. The checklists used in the OSCE presented here were carefully constructed according to the principle that each item represents a specific unitary behavior that is easily observed to be performed or not performed. This approach is akin to Clauser’s idea of analytic scoring. This quality greatly reduces error in performance assessment checklists and is essential when the raters do not have extensive training with the checklists (Clauser, 2000).

A majority of our OSCE faculty has experience precepting OSCE’s at our school. All received training for this OSCE, as discussed in the last sentence of p.4 in the originally submitted manuscript: “All preceptors received standardized guidelines for checklists and feedback prior to each OSCE session.” Since our OSCE’s were all scored in a similar fashion under similar types of setups, the precepting skill is highly transferable. Based on Dr. Mavis’ concerns regarding rater training and checklist use, we have extended our description of these points in the Methods at the top of p. 5, second and third lines and last sentence of the first full paragraph, respectively (Attachment, A.1c.).

Finally, OSCE checklists in this study are being used to rate second-year medical students and not experts. Checklists do not perform well when experts are being rated (Hodges et al., 1999), since experts use heuristics which second-year medical students are not likely to employ (Clauser, 2000; Maudsley, 2000). These undergraduate medical students are therefore better measured by checklists, which seek to document each step involved in the performance of clinical tasks.

Our response to Dr. Mavis’ second major concern follows.

2. [whether] the stability of the OSCE over time is as much an artifact of rater bias obscuring real differences in performance. …and [are the results of the OSCE] used to modify instruction with an effort of improving educational outcomes?

Dr. Mavis quite rightly calls into question whether or not the stability of the OSCE over time is a true measure of student performance or is an artifact of rater bias. However, IRR would not really address this issue. In order to increase IRR, test items will often become over-homogenized (Fan & Chen, 2000). These items are easy for raters to agree on, but often at the price of not really measuring the full range of student performance.

In this critique, Dr. Mavis alerts us to the question of test validity – i.e., is the instrument really measuring what it is supposed to be measuring. This potential problem for all OSCE’s is usually approached the same way that validity is determined for other types of test instruments. As Clauser (2000) puts it, “…validity arguments take the form
of a chain of evidence linking a performance to a target domain through a score.” (p. 311).

With regard to our OSCE, we have some indicators of validity, which lead us to believe that we are actually getting good measures of student performance and not rater bias. We described some of our validation in the Methods of our original submission, under “Objective Structured Clinical Examination (OSCE): Development,” at the end of the second paragraph on p. 4. “Site directors determined that all OSCE questions reflected essential skills to be mastered by second-year students. … Annual feedback from students and faculty endorsed the face validity of the OSCE. In 1999, 90% of students and 91% of faculty agreed that the OSCE represented an appropriate and fair evaluation method.

More recent work from our group extends the validation of our OSCE. In a study examining congruent validity presented earlier this year at Society for General Internal Medicine (Simon SR, Volkan K, Hamann C, Duffey C, Fletcher SW, 2001). We found a strong relationship among student ability upon matriculation to medical school, first- and second-year course performance, and OSCE scores. Other studies have reported similar results (Merrick, et al., 2000). We now discuss validity in a separate paragraph of the Discussion, p. 9, paragraph 4 (Attachment, A.2.).

Another manuscript in preparation from our medical school shows a strong relationship between checklist scores and global ratings of students on the OSCE. An internal, unpublished study of the internal validity of the OSCE demonstrated that a latent trait model of the OSCE showing two latent factors related to student knowledge and performance ability was a good fit to the data. Finally, data from a similarly structured 3rd year OSCE at our institution showed congruent validity by explicating the linear and non-linear relationships between clerkship grades, time from clerkship and OSCE scores in a radiology station (Morag, Lieberman, Volkan, Lang, 2001).

Though we fully acknowledge Dr. Mavis’ two major concerns, we believe that our OSCE and these studies are useful and support the reliability and validity of our OSCE, especially when seen in the light of the stability of the scores. Nevertheless, we plan future studies to compare the generalizability and rater-reliability approaches.

We now discuss Dr. Mavis’ first concern.

3. The research questions/purpose of the study are not clear…and [the study] does not appear very generalizable to other settings.

We agree, have clarified the research questions further in the Background, p.3, paragraph 4 (Attachment, A.3.a.), and added comments regarding generalizability to the Discussion, p. 10, new paragraph 3 (Attachment, A.3b.).

The OSCE has been useful to our faculty for curriculum planning. The data from the OSCE have also been useful to our second-year students and the site directors for our physical diagnosis course, especially to help identify areas of weakness that could benefit
from remediation prior to the start of clerkships. This benefit is especially true for students with the lowest performance on individual stations and skills (see Discussion, p. 10, paragraph 3).

The scientific literature on OSCE is in an “inductive science” phase, and many formative studies of OSCE are being published. Reports of OSCE’s developed at a single medical school have great usefulness for the field because they allow researchers and educators to adapt the best published practices to their specific medical school (van der Vleuten & Swanson, 1990). Furthermore, there are few reports of OSCE’s at the level of training equivalent to the second year in the U.S.A. and no comprehensive ones, as we stated in the Background of the original submission, p. 3, paragraphs 2 and 3. The utility of our methods, as described in our responses 1 and 2 above, and of our OSCE should extend to all medical schools that face the logistical challenges posed by multiple sites and preceptors for student training in physical diagnosis.

References for Response A:
Fan X, Chen M: Published studies of inter-rater reliability often overestimate reliability: Computing the correct coefficient. *Educational & Psychological Measurement* 2000, 60: 532-542
Longford NT: Models for uncertainty in educational testing. *New York: Springer-Verlag* 1995
B. Response to review by Dr. Townsend

1. …the running title is misleading… We agree, and have changed the running title (header) to “OSCE Assessment of Physical Diagnosis Skills.”

2. …some details of the questions used in the stations would be helpful… The substantial number of differential diagnosis questions (30) concerning calf pain was derived as follows: 6 differential diagnoses, and a range of 1 to 7 questions concerning physical findings characteristic of each of the 6 differential diagnoses. Space considerations precluded our inclusion of this level of detail from the overview provided in Table 1.

3. …no comment has been made about whether or not students were able to complete the tasks in the allotted time… We have clarified this point with additional text in the Methods, p. 4, last two lines of the second full paragraph.

Attachment

Underlined text changes as found in revised Manuscript

“How Well Do Second-Year Students Learn Physical Diagnosis? Observational Study of an Objective Structured Clinical Examination (OSCE)”

A. Response to review by Dr. Brian Mavis

1.

a. Methods, top of p. 6

Because it was not logistically possible to obtain inter-rater reliability due to the large number of preceptors, we used generalizability theory analysis [27]. This analysis accounts statistically for rater error by parsing out the variance relevant to the instrument in question. By modeling the variances as separate characteristics, we isolated the variance due to student ability, which in classical test theory is equivalent to true score variance. Other variances related to the test are treated as error variances. In this framework, we treated error due to differences in raters as error variance.

b. Discussion, starting at the top of p. 9

This study has several limitations. We have not directly assessed inter-rater reliability because of logistical and cost constrains. To address methodological concerns, we used generalizability theory (GT) to produce a measure of reliability similar in quality to inter-rater reliability [31].

There are a number of examples of the use of GT to account statistically for rater error [30,32-34]. Using GT can also overcome some problems inherent in inter-rater reliability, such as overestimating reliability [35]. Due to the large number of preceptors involved in our OSCE, we made the statistically reasonable assumption that any error due to rater differences is randomly distributed. Since randomly distributed error has a mean of zero, the error variance due to differences among all preceptors is zero. In our OSCE, the variation of individual raters around the mean station score of all raters is very close to 0 (e.g., .04 for the presentation station, data not shown) and the standard deviations of student scores are comparatively large (e.g., 15 for the presentation station). Finally, our
GT-based assumption is especially appropriate when the test scores used in the analysis are created by summing many items across each scale. Summing in this fashion has the effect of further randomizing the error variance. The reliability, or internal consistency, of the overall OSCE was good at .86. The reliability of 6 of 7 skill scores, and 9 of 16 station scores, were acceptable at > .60.

Another benefit of the GT approach is that the reliability coefficient derived from the GT analysis is equivalent to Cronbach’s alpha coefficient which, for binary items, is equivalent to the KR-20 reliability coefficient. The alpha coefficient is especially useful during test development because it gives a measure of how each item is contributing to the scale to which it has been assigned. This measure makes it easy to target items for editing or deletion if they are not performing well. Since we are ultimately interested in using the scale scores for our research study, the GT measure of reliability is appropriate for OSCE’s involving many preceptors.

c. Methods, top of p. 5

[All preceptors] received standardized guidelines for checklists and feedback prior to each OSCE session, as did the standardized patients or actors for the abdominal pain, alcohol/abdominal exam, knee and thyroid stations. Fourteen stations were each 6 minutes in duration, and two - abdominal pain and headache - were 12 minutes in duration.

At each station, the student performed the indicated tasks […] Each year, approximately 150 preceptors participated in the OSCE, and 60% have had experience with this OSCE and the checklists from prior years.

2.
Discussion, p. 9, paragraph 3

The validity of our OSCE is only partially established. […]. Finally, a more recent investigation supports predictive validity of our OSCE. Physical diagnosis skills examined in the present study correlated with scores on the USMLE Step 1 exam, and the skills that foreshadow the clinical clerkships - identification of abnormality and development of differential diagnoses – best predicted USMLE scores. [41]

3.
a. Background, paragraph 4

We used the OSCE to examine how well second-year students learned clinical skills in the second-year physical diagnosis course at Harvard Medical School. We were particularly interested which skills students performed best and which were most difficult. We assessed what factors affected their performance on the overall OSCE, and on individual skills and stations. Finally, we examined whether student OSCE scores varied from year to year, medical students performed differently from dental students, learning at different teaching sites affected student performance, and preceptors and examination logistics affected student scores.

b. Discussion, p. 10, paragraph 3

The successful implementation of the OSCE at our medical school is likely relevant to all medical schools that face the logistical challenges posed by multiple sites and preceptors for student training in physical diagnosis. Furthermore, the results from
the second-year reported here and our pre-fourth year [Morag] OSCE’s have been useful in helping to identify areas of weakness that could benefit from remediation prior to the start of clinical clerkships. This benefit is especially true for students with the lowest performance on individual stations and skills. For site directors and faculty, the OSCE has also helped identify those parts of the curriculum students had difficulty mastering. Identifying opportunities for remediation through a second-year OSCE held prior to the end of a physical diagnosis course allows medical school faculty to structure the remaining sessions of the course and to improve future physical diagnosis teaching.

B. Response to review by Dr. Townsend

Comment 3. p.4, second full paragraph, last two lines

In 1999, 90% of students and 91% of faculty agreed that the OSCE represented an appropriate and fair evaluation method, and that enough time was given to complete the stations.

REFERENCES

We included 8 additional references:


38 Longford NT: Models for uncertainty in educational testing. *New York: Springer-Verlag* 1995


40 Fan X, Chen M: Published studies of inter-rater reliability often overestimate reliability: Computing the correct coefficient. *Educational & Psychological Measurement* 2000, 60: 532-542
