Reviewer's report

Title: Temporal Stability of Objective Structured Clinical Exams: A Longitudinal Study Employing Item Response Theory

Version: 1 Date: 23 September 2012

Reviewer: William Roberts

Reviewer's report:

The purpose of this study was to assess the psychometric fidelity of clinical skills performance scored in the same six OSCE stations at two different time points. The cross-sectional comparison of candidates and raters included testing that took place in 2007 and 2008. Generalizability theory and the 2-parameter logistic model were applied for the analysis. Results indicated stable scored performance in each of the exposed six OSCE stations over time.

Major Compulsory Revisions

This review includes a list of comments to address areas of strength and weakness with recommended corrections. Technical and general areas of weakness in the manuscript will need to be addressed before acceptance for publication.

1. Specify in design notation the effects for the single facet nested G-study design. Include if this is a norm referenced or criterion based examination. If this is a criterion based examination, then the index of dependability is the appropriate measure of reliability. Report standard errors and variance associated with each facet.

2. Report more specifically the psychometric indices that are mentioned at the end of the first paragraph under the Procedures subheading.

3. It is not clear if reliability of scores was assessed from performance across all 10 stations or only on the subset of 6 that were repeated at time 2.

4. A number of clinical skill measures are mentioned. Exactly what measures were included for this study? Describing measures in their own subheading would be useful. Is the examinee’s score the sum or percentage of correct checklist items across all 10 stations or just the 6 repeated stations?

5. Clarify if the standardized patients and physician rater are confounded; that is, their effects can or cannot be separated in the G-study.

6. Describe the clinical case content for each of the 6 OSCEs with repeated material.

7. Include IRT standard error of estimation for each station.
8. Scale values are missing for both ordinate axes in Figure 1. Should these axes be labeled probability rather than proportion correct?

Minor Essential Revisions

1. The 2-PL and 3-PL models are presented in the paper, but there is no mention of the 1-PL Rasch model. The Multifacet Rasch model has been used in several studies to account for the effects of rater stringency, case difficulty and other facets typical to OSCE examinations. I would prefer to have seen some discussion about the consideration of these models for analyses.

2. I am concerned that principal component factor analysis was compared between the two times of test administration for two of two stations without more discussion about these results. This analysis seems to be an attempt to test factor structural invariance between the two test administrations. Test of invariance over time is possible with confirmatory factor analysis. In this instance, the analysis is exploratory that could be investigated in the future with confirmatory factor analysis.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.