Reviewer's report

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

Version: 1 Date: 11 October 2012

Reviewer: Celia Taylor

Reviewer's report:

Major compulsory revisions
1. The introduction is somewhat repetitive and jumps around from one issue to another and therefore could be better structured and shortened.
2. The introduction has not convinced me as to why stability over time is so critical – particularly when there are only two years’ worth of data. I would be more concerned about the performance of new stations. This partly depends on the approach to standard setting taken for this OSCE, e.g. the same Angoff used in both years vs. BLG/BLR undertaken each year – but details on standard setting are not provided.
3. Examiners only received 30 minutes of training – this does not seem very much and thus the sufficiency (or otherwise) of the training should be considered in the discussion.
4. In an IRT model the mean item difficulty across all items (stations) should be 0; yet all stations have a negative item difficulty. This result needs to be explained.
5. The authors conclude that non-significant results mean that the stations are stable but the results actually mean that a difference has not been detected. Without a power calculation it is not possible to determine the reliability of this interpretation.

Minor essential revisions
1. Correction of typos and grammar errors and inconsistencies (e.g. 1980s vs. 1980’s; use of numerals for numbers <10).
2. The introduction needs more clarity with respect to what is meant be ‘stability’.
3. In the methods it is noted that “each year the AIMG program re-uses 4-6 OSCE stations from previous years with good psychometric indices”. The authors need to specify what these psychometric indices are and what makes them “good”.
4. The formulas need to be explained further, e.g. what is varcomp? Where is Cg in the IRT model formula? Is the IRT model used at checklist item level or at station level?
5. The use of a simplified G model needs to be explained and its limitations discussed.
6. The G coefficients are presented in the methods section but are actually results. What is meant by varying stations and why has the model not been used to evaluate overall generalizability? It is noted in the discussion that the G coefficients were “acceptable” yet the definition of “acceptable” is not defined in the paper.

7. In the methods section the authors need to explain why they have used MANOVA (e.g. what null hypothesis they are trying to test).

8. The use of exploratory factor analysis needs to be explained further. I had to assume that the TCCs computed with homogenous items used the results of the factor analysis. However I am not convinced what this approach (cf. using overall TCCs) adds to the paper. Again the sentence “The pattern of TCC for homogenous and heterogeneous items for all the OSCE stations was similar” belongs in the results section, not the methods section. If the factor analysis is retained, the full results should be shown in an Appendix. The third finding summarised in the discussion relates to the TCCs and presumably is based in a comparison of the TCCs in Figures 1 & 2 – this needs to be explained in the text. When this issue is considered again later in the discussion I think the word heterogeneous has been used instead of homogenous – without clarity here I struggled to understand the comments on clinical skills being contextual.

9. In Table 1, explain what is meant by “Means” in the column headings for Time 1 and Time 2. The full test statistics and exact p-values for all tests should be provided. The authors need to specify how they have corrected their results for multiple comparisons.

10. Figure 1: the y-axis needs labelling. Figure 2: x-axis scales have different centre points. It is a little confusing that the proportion correct = difficulty, since the term ‘difficulty’ has a specific meaning in IRT.

Discretionary revisions

1. The details of the Time 1 and Time 2 OSCEs are not very clear to a reader unfamiliar with the system and hence this section could be re-worded to improve readability.

2. Time 1 and Time 2 are used in the text but 2007 and 2008 on the figures. I suggest one or the other is used throughout.

3. I do not understand what is meant by “the slope of the curve” in the sentence “the overall discrimination (the slope of the tangent at the break in the curve” (sic) and this should be clarified.

4. Descriptions of the 6 stations would be helpful, but may not be provided for reasons of security.

**Level of interest:** An article of limited interest

**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.