Reviewer’s report

Title: Reliability of Multiple Mini-Interviews and traditional interviews within and between institutions: a study of five California medical schools

Version: 0 Date: 15 Jun 2017

Reviewer: Timothy Wood

Reviewer's report:

Thank you for the opportunity to review the paper titled "Reliability of multiple mini-interviews and traditional interviews within and between institutions: a study of five California medical schools". I found the idea of comparing interview and MMI measures for the same students at five different schools to be intriguing, and this paper should be of interest to people involved in admissions. I have a number of concerns though which if addressed with strengthen the manuscript.

The introduction (Page 4 to 5) uses a comparison of reliability between measures as the driver behind the study. To me this type of comparison is of limited value. Any two measures can have extremely high reliabilities (or low) but unless the content, the examinees and perhaps even the raters are identical, making comparisons of reliability is of limited value. What is important is the degree the measures provide the same scores and the degree that decisions made from those measures the same. Note that I am not saying that reliability should be ignored, it is important because the degree of unreliability (i.e., measurement error) will attenuate any comparison between measures. My suggestion is to de-emphasize the comparisons of reliability between schools as a driver of this study and emphasize the comparisons between scores and acceptance decision. This is the unique aspect of what they have done and so should be emphasized.

I have to admit, I found the analysis and results section to be unclear. First, the authors ran a generalizability analysis to derive an estimate of reliability. It looks like the measures (T1-T3 and MM1, MM2) were nested within person. I think their model is not as well specified as it could be because I should not have to assume what the model looks liked. I also don't think nesting the measures within person is appropriate. To me the measures are crossed, there are people who have scores on all five measures. By nesting, the assumption is that the measures associated with person 1 are different than the measures associated with person 2 but that is not true. My other concern is with year. The people and even the examinations used in 2011 will differ from 2012 and 2013 but year is not a factor in the model or in any other analysis for that matter. To that end, I would likely have people nested within year and crossed with measure in order to include year explicitly. Or alternatively, analyze each year separately in a person x measure design. Finally, the results of the generalizability analysis are not reported according to best practice guidelines. See Crossley, Russel, Jolly, Rickets, Roberts, Schuwirth & Norcini (2007) for advice on how to report a g-study.
My second concern with the analysis and results is that there is a line in the within school comparison that says that the interviewer ratings for TI1, TI2, and TI3, as measured by a Pearson r correlation were lower than the Cronbach's alpha for the MMI and MM2. This is a false comparison. For starters the TI measures that were reported are comparisons between raters whereas the MMI measures are comparisons between stations on the examination so comparing them is questionable. Second, a Pearson r ranges from -1 to +1 whereas Cronbach's alpha ranges from 0 to 1 so they are on different scales. This section needs to be revised. I am also unclear why a correlation between raters was used as the measure of reliability for the TI's when the manuscript states that interviews all used standardized scales. If that is the case then a Cronbach's alpha or even a g-coefficient could have been reported for an estimate of reliability for the TIs.

Table 3 is of primary interest but it does needs some clarification. First off, year is not a factor in this table. The people and content in each year is different and therefore I do not think that the authors can generate correlations without factoring that in somehow. My second concern is that in addition to observed correlations, I think the authors would benefit from reporting dissattenuated correlations. A low correlation can occur because to measures are legitimately unrelated or because the amount of error in the measure (i.e., unreliability) will attenuate the correlation. By deriving dissattenuated correlations, one can see what a correlation would have been if the measures had been perfectly reliable. If a dissattenuated correlation is high (> .85) then one knows that a low observed correlation is due to poor reliability. If a dissattenuated correlation is low to moderate then one knows that a low observed correlation occurred because the measures really are unrelated.

In addition to the correlation between scores, I think it would have been interesting to see to what degree the acceptance decisions agreed between measures.

Page 10 states that the reliability for MMI2 is lower than MMI1. How was that determined since there was no statistical test to compare the values?

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

No
Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons
CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal