Reviewer's report

Title: Evaluating the validity of an integrity-based Situational Judgement Test for medical school admissions

Version: 2  Date: 29 January 2015

Reviewer: Kulamakan Kulasegaram

Reviewer's report:

Generally, this is a very interesting paper. I have no major issues with the design or premise of the study. This is by and large how most admissions research is conducted and proceeds on the basis of predictive and concurrent correlation between measures in order to improve information on candidates.

I think there are a few issues though that must be addressed first.

Major

1. The authors claim the we 'know' what the MMI measures. Actually, the data are quite contradictory with some studies showing low-moderate personality factor correlations, a recent CFA pointing to communication and some other large factor (Oliver Adv Health Sci Educ Theory Pract. 2014 Aug;19(3):379-92) and there is lots of disagreement Knorr M Med Educ. 2014 Dec;48(12):1157-75.). This is mostly because the MMI is a method, not necessarily a fixed tool. What you put in is what you get out.

In light of this, are the correlation coefficients interpretable? Can the constructs of the SJT vs. specific MMI stations be addressed? Low reliability of a single station can be a problem but see the point below for a solution.

2. The authors have ignored a very large set of studies by Dore et al. on the use of SJTs with predictive validity data now available and showing high correlations with MMIs. These SJTs are open-ended and video based but they are essentially the SJTs that Lievens most recently called for (January 2015). Adding this data and discussing the lack of correlation is going to be useful for interpreting the generalizability of the results.

I am not saying that this means that the authors' SJT form is weaker than Dore's but rather, each SJT was geared toward specific constructs with Dore's being wider than the authors.

3. The low correlations - both restriction of range and low reliability of the measures should be corrected for. The formulas can be found in Health measurement scales 3rd edition (Norman and Streiner) or other psychometrics resources. Correcting for these can indicate the true value of the correlation and help interpret the results. Analyzing specific stations against the SJT scores after reliability correct can be a useful Hypothesis generation exercise (and obviously...
not appropriate for a primary outcome).

4. The authors should report more data on the response patterns for SJTs. They do so with the removal of items which reduced cronbach's alpha but the report could benefit from indices of difficulty. If possible, the authors could use a polytomous rasch model to generate average indices. If not, the classical test theory versions of these statistics should at least be presented.

Lastly, it seems in general, low correlations (<0.4) between personality factors and everything else seems to be the rule. Could it be possible that these tests are far too transparent for utility in selection?

Otherwise, great paper, enjoyed reading it.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

No completing interests to all questions.