Reviewer's report

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

Version: 2  Date: 5 February 2015

Reviewer: Mark Albanese

Reviewer's report:

This is an extremely interesting and potentially important study. SJT and integrity seems quite compatible. The approach used for generating the instrument is generally quite impressive. The potential for having a written instrument to assess integrity could be very helpful. That said, there are a number of issues.

Major Compulsory

1. The method used to recruit participants into the SJT treatment needs to be described in greater detail, particularly if there were any incentives provided.
2. The authors merged data from 3 types of formats: written, video and verbatim transcripts on the strength that the results were “broadly similar”. More details are needed before such a merging can be accepted.
3. As a new instrument, it would be helpful to put it through the paces for determining construct validity. A factor analysis is one of the typical construct validity assessments.
4. The study sample is actually somewhat biased based upon the differences in MMI results for the SJT versus everybody. This should be noted in the discussion and some consideration as to the generalizability of the results given.
5. The degrees of freedom for the t-test are confusing. There were a total of 480 subjects who were given the MMI. How can you have from 547 to 551 degrees of freedom? Where did the extra 70 or so subjects come from?
6. It is not clear how many scorable units there were in the SJT associated with each of the 10 scenarios as well as overall. Given that the questions associated with a given scenario are more likely to have a higher correlation with each other than with questions associated with different scenarios, reliability estimation becomes complex, as does the scoring. More attention needs to be given to this.
7. On page 13, the SJT group has lower MMI scores than the non-volunteers, however, it appears to be a difference in means that are less than the .2 SD values considered a small effect size. The effect size is a more accepted index of differences than is the percent of total that is currently presented.
8. The score scale for the SJT is confusing. In text, the mean for males and females was given as 4.58 and 5.93, respectively (p. 14). In table 3, the overall SJT mean is 31.58 with an SD=7.15. Please indicate what the score scale is for the SJT and be consistent in both the tables and text.
9. If the males do perform significantly worse on the SJT, as indicated in text,
there is a bias that needs to be taken into account in computation of both the correlations and reliability estimates. Otherwise, the small correlations noted could be due in large to the group differences.

10. In and of itself, the gender difference has an effect size on the order of .46, approaching a medium effect size by Cohen’s criteria. This is a significant problem for any use of the instrument and should be addressed as a major limitation that needs further study.

Minor essential
1. Please define what a “Scottish Higher” (p. 13) happens to be. Some readers may not know what an A-level is either.

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:
I declare that I have no competing interests