Reviewer's report

**Title:** Analysing clinical reasoning characteristics using a combined methods approach.

**Version:** 5  **Date:** 23 September 2013

**Reviewer:** Mike Tweed

**Reviewer's report:**

The issue of investigating the information provided by different methods of assessment that may complement each other is worthy of investigation and reporting. The manuscript has been improved significantly and is more straightforward to follow the rationale, methods, results, and discussion.

**Major revisions**

The major issue relates to the purpose, which has been clarified as testing using a combination of SCT and CRP to assess clinical reasoning characteristics so that potentially medical student learning can be customised. The results and discussion should focus on this.

**Introduction**

End Para 2 and 3: As no single method has been shown to be reliable and valid for all purposes and contexts, the authors quite rightly state that there is no gold standard. However, they finish para 3 with the statement regarding CRPs and SCT demonstrating reliability and validity across a variety of contexts. If individual methods have demonstrated validity for some purposes in some context then, as the authors point out and have done, combining complementary methods may be valuable. Hence this study.

**Method data analysis:**

If using correlations to provide evidence for construct validity, a theoretical basis as to why SCT and CRP F score correlation would support this needs to be included. Then a lack of correlation with d-score should be added, as validation by showing that correlations that should be lower, or non-existent, are lower or non-existent, adds to the evidence.

The “pass” mark might be better phrased. Was the mark meant to be equivalent to a standard, so that all candidates would know where they were in relation to that standard, or as a mark that might indicate those candidates scoring lower are likely to be at greatest need of a change of learning strategy? This mark does need to be linked to the purpose. “Pass” might not be the appropriate term.

Was the “pass” mark derived from all scores, or just the medical students? I don’t think it was not pass-mark by cohort.

**Results**
As pointed out there was incomplete data for all candidates over all their questions. Where this was possible all questions answered—were the results any different?

If I understand correctly candidates, who were either year 3, year 4 medical students or GPs, were allocated to sit SCT1CRP2 or SCT2CRP1. Presenting the results as SCT1CRP2 and SCT2CRP1 might re-enforce this.

The abstract mentions correlation of CRP2 and SCT2—but I thought that the same candidates did not sit these two questions.

Where multiple analyses have taken place, is allowance made for this?

Assuming that there was one “pass” mark for all, rather than a pass mark by cohort, how many GPs did or did not achieve the pass-mark?

Discussion

Para 1: The authors raise the possibility that more able students may have been disproportionately represented, but also highlight the low pass rate across the methods. These cannot both be true, unless even the more able are still not suitable to pass, which I doubt. It may relate to the choice of the term “pass-mark”.

Para 2: Although increasing number of problems would increase reliability, increasing the diversity of the candidates may also do this. It could be that the low number of GP responses and homogenous student ability (as noted in para 1 of discussion) was a factor.

Why did the SCT not demonstrate any difference between the groups? If the SCT cannot discriminate between members of groups who are assumed to have different levels of clinical reasoning, how can it be used to guide learning for that characteristic? This is worthy of discussion.

Conclusion

This has to relate to the purpose. Combining test results to improve feedback to guide learning is definitely worthy or study and reporting. The statement that these methods in combination might be used for teaching rather than assessment does leave me wondering if they should be abandoned to guide learning. I hope not.

The past sentence talks of a “two-stage” process. This has not really been investigated and so replacing that with “two-method” process might be better.

Discretionary Revisions

Introduction

Para 6: “Acceptable” reliability depends on the purpose of the assessment. Less reliability may be appropriate if the purpose is guiding learning rather than making progress decisions.
Para 6, Last but 1 sentence: could start “However, the SCT cases…..

Method: procedure
Sentence 1: To make the process easier to understand. Complementary might be replaced by opposite (or similar).

**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.