Reviewer’s report

Title: Selecting Optimal Screening Items for Delirium: An Application of Item Response Theory

Version: 2 Date: 14 August 2012

Reviewer: Edward Ip

Reviewer’s report:

This is a well written paper with extensive and careful work on the application of novel methodology to the selection of items for a screening tool for delirium. Tools from modern test theory, namely the IRT, is suitable the stated purpose. The sample size for calibration and for the development of the screening tool adds to the strength of the paper. Thus, the work appears to have important clinical impact in terms of its application to the clinical assessment of delirium.

Despite of the strengths the work does not appear to be highly innovative in terms of its methodological contribution. The paper applies existing methods to a domain area. As the paper appears to be a revised resubmission, I would relegate the decision regarding how the level of innovativeness in methodology is handled to the editor and instead focus on providing specific comments. If innovation in methodology is not a primary concern, then I think the paper passes the threshold necessary for publication.

Specific comments (Minor essential revisions):
1. p.3 line 5. line 74: punctuation problem and incorrect capitalization.
2. p.6 line 105 and 106. I suspect that the two references need to be interchanged.
3. p.6 line 109. It's better to characterize IRT as an extension to classical factor analysis, which is what Mislevy's paper did.
4. p.7 line 119. y{i} is response and does not belong to the parameter set. I'd also suggest using the term person-level trait or similar terms for theta; it is not universally agreed that theta can be described as a parameter especially within the marginal IRT models.
5. p.7 line 124. It may help the readers in understanding the meaning of theta in a clinical context by stating its interpretation for this specific application. Additionally, it is important to make it clear that higher theta means higher propensity for delirium.
6. p.8 line 149 - p.9 line 185. The discussion for distinguishing between the two kinds of tests could be somewhat streamlined. It's a bit excessive to use almost 2 pages of text and a diagram to explain that. Further, the distinction did not seem to get referred to later (see comments #14 below).
7. p.9 line 186. I was a little surprised that the substantial literature for developing
item pools, short forms, and adaptive tests from the very recent PROMIS activities has not been cited/mentioned at all. Guidelines for item screening, for example, from Reeves et al (2007, Med Care) could be used here.

8. p.11. Measurement. A tally of the items being considered from the source instruments is useful.

9. p.11. Measurement. It is not clear what types of response categories (dichotomous/polytomous/categorical) were present in the source instruments. The equation in line 124 is only applicable to dichotomous items, so clarification about how potentially different response categories were handled is needed. If polytomous IRT response model was used for item calibration, it also needs to be clearly stated.

10. p.13 line 268 check for multicollinearity. Is it simply cross-tab or local dependency (i.e., correlation given factor)? A brief explanation is needed.

11. p.15 line 311. Typo at the first position.

12. p.15 line 312. The dimensionality assessment is rather elaborate. I assume that when the m subsets were used to achieve unidimensionality, a score for each dimension would be developed. If I'm following the logic correctly, then from Table 2 (looking ahead), there will be a total of 10 dimensions from the Feature Sets. For the stated purpose of screening, how will the 10 scores be used? Will a composite be developed?

13. p.18 line 390. ... one such curve.... the last sentence talks about item information at the 50th percentile which is a number. This sentence needs to be reworked. Perhaps the word "curve" could be added after information in the previous sentence.

14. p.19 line 409. Tying it back to the predictive use of the tool (Figure 1) could help readers understand the features here.

15. p.22. line 468. I believe that perhaps the most important future work is to establish the validation of the instrument for the stated purpose. No statistic has been presented on the predictive validity (e.g., classification error for CAM delirium) in this paper.

**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