Author's response to reviews

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

**Authors:**

Frances M Yang  (francesyang@hsl.harvard.edu)  
Richard N Jones  (jones@hsl.harvard.edu)  
Sharon K Inouye  (sharoninouye@hsl.harvard.edu)  
Douglas Tommet  (dougtommet@hsl.harvard.edu)  
Paul K Crane  (pcrane@u.washington.edu)  
James L Rudolph  (jrudolph@partners.org)  
Long H Ngo  (lngo@bidmc.harvard.edu)  
Edward R Marcantonio  (Emarcant@bidmc.harvard.edu)

**Version:** 3  **Date:** 26 October 2012

**Author's response to reviews:** see over
Dr. Monica Taljaard  
Editor: BMC Medical Research Methodology  
Epidemiology & Community Medicine  
451 Smyth Road, Room 3105  
Ottawa, Ontario, Canada, K1H 8M5  

RE: Manuscript 5082650927441217 - Selecting Optimal Screening Items for Delirium: An Application of Item Response Theory

Dear Dr. Taljaard:

Thank you very much for the reviews of our above-named manuscript. My co-authors and I have reviewed the comments and made revisions to the manuscript, which is attached to this note. In this letter we discuss how we have addressed each of the reviewers' concerns.

Reviewer: 1

Reviewer's report  
Title: Selecting Optimal Screening Items for Delirium: An Application of Item Response Theory  
Version: 2 Date: 13 August 2012  
Reviewer: Claudia Spies

Reviewer's report:

Yang and Colleagues performed a complex and very promising analysis of a database of screening data from a multicenter trial in order to identify indicators for delirium according to the CAM features. Generally speaking this study is of high interest for delirium research; the methodology is sophisticated but very difficult to follow. For readers unfamiliar with statistical methods such as ‘permuted parallel analysis’ (page 14, line 281) the results e.g. presented in Table 2 of the manuscript are very hard to follow. The overall analysis presented by the authors is a mixture of theory- and data-driven analyses and expert consensus decisions (e.g. lines 303-304: “We considered the preponderance of the evidence in making dimensionality decisions, together with the input of the CEP on the interpretation of secondary factors.”). It remains unclear from the description whether this approach is common praxis or was developed by the authors. If it was common praxis authors should provide the rational for this. We would suggest
asking a separate statistical review by a statistician familiar with factor analyses and item response theory to judge the appropriateness of the statistical methods used.

Response: The work we summarize in this manuscript is consistent with standard praxis. However, the use of item response theory (IRT) is an innovative method for developing a delirium screening tool that has not yet been applied for the best of our knowledge in delirium research. We have made some general revisions to the methodological presentation to improve clarity and accessibility. These revisions are described more fully in our response to comments made by Reviewer 3.

It remains further unclear how the authors considered potential differences in age and gender. One may imagine that some of the source and indicator items (lines 235-237) were gender-specific or age-specific and it is unclear if this was considered by the authors. E.g. Jane and colleagues identified a gender bias in diagnostic criteria for personality disorders in DSM-IV criteria (Jane JS, Olmanns TF, South SC, Turkheimer E. Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. J Abnorm Psychol. 2007 Feb;116(1):166-75). In the actual dataset more than 2/3 of the participants were women and it remains unclear whether this fact was adequately considered by the authors.

Response: We did not consider differences in the measurement of symptoms of delirium between men and women and persons of different age groups. We appreciate that the reviewer has identified gender and age differences in personality disorders, and we are aware of gender differences in other psychiatric and cognitive disorders. This is a reasonable research question, because there is some previous research showing differences in the expression of some delirium symptoms with age, although we are unaware of any research showing sex differences in the expression of symptoms of delirium. Insofar as the application of item response theory methods to the symptoms of delirium is entirely novel, we think it is best to reserve advanced applications, like differential item functioning detection, for future research. In response to this concern, we have added to the limitation section that factors related to the expression of symptoms by age and gender have not been considered in this analysis, and that future analysis may be warranted to examine measurement bias in the instrument.

To address the reviewer’s concern, we have added the following to the text:

(p. 23) Another limitation of our analysis is that age, sex and race/ethnicity, have not been considered in this analysis. These factors have been shown to be associated with the differential expression of signs and symptoms in other psychiatric and cognitive disorders, although not necessarily in delirium. Our results assume that the measurement of symptoms of CAM features is invariant across major sociodemographic groups. A future direction for potentially improving the current instrument is to examine measurement bias due to age and gender.

From a clinical point of view a database with more than 4500 cases does seem appropriate for the overall task of this study. However, there are some aspects worth noting.

Major comments:
Was Delirium determined according to DSM 4 criteria? Or according to ICD 10 criteria? Both are the current “Gold Standards”. If the CAM was used as a reference standard does this not exclude part of the delirious patients? Or are DSM 4 and ICD 10 and CAM equal? Please explain in detail and add in detail to the limitations section.

Response: We appreciate the opportunity to clarify our goals and methods. This study was not intended to propose CAM as a reference standard, nor to compare it with either DSM-IV or ICD-10 criteria. Given that the CAM is widely used in the literature and is recommended for bedside screening\(^4,5\), our goal was to identify the most efficient set of items that non-experts can use to determine the presence or absence of the 4 CAM diagnostic features.

We have made revisions to better explain our method:

(p. 8) The goal of this paper was to identify the shortest set of mental status assessment questions and interviewer observations that could be used to efficiently provide relevant information for screening about a patient’s level on four CAM diagnostic features.

In addition, we note in the limitations section the DSM-IV and ICD-10 were not used as diagnostic criteria for delirium. We now emphasize this point with the following sentence:

(p. 23) The DSM-IV and ICD-10 are used for diagnosis and coding by trained clinicians. In contrast, the design and purpose of the current study was to identify items for delirium screening based on the four CAM features, which can be done by both clinicians and trained non-clinicians. Therefore, this research may not directly inform diagnosis relying only and strictly on the DSM and ICD.

Subsyndromal / pre-delirium as well as severity of delirium is an issue, IRT does seem to be a suitable method to encompass a less simplistic definition of delirium than the CAM is offering. Please explain how IRT can be used to define Delirium in an ordinal approach.

Response: The issue of continua and categories in psychiatric diagnosis is a longstanding point of concern in psychiatric diagnosis.\(^6\) IRT and latent trait models have an important role to play in this area of research and development,\(^7\) but the appropriate models are hybrid models that simultaneously include latent traits (continuous dimensions of pathological thought or behavior) and latent categories (level of impairment or other qualitative distinctions from “disordered” and “non-disordered”).\(^8\) Because we thought this discussion was beyond the scope of this manuscript, we did not change the manuscript in response to this comment.

Delirium is a complex and multifaceted disorder. In our manuscript we only try to convey an approach to identifying a reduced item set for most efficiently assessing the presence or absence of four key features of delirium, reflecting at least four dimensions of potential impairment. Each dimension is a latent trait, and describes a continuous dimension upon which thresholds demarcating “normal” from “abnormal ” levels of impairment can be identified. Finding those thresholds is not something we endeavored to do and represents a logical progression but a separate line of research from what we have presented.
Of the 4744 Patients, how many displayed Delirium? How was the consent handled in delirious Patients? P.11 I.224 How was the consent handled in Patients with low MMSE values? This needs to be explained

Response: Of the 4593 patients that were included in the analysis, 13% displayed CAM delirium, which represents 611 patients. Patients were only included if they were able to provide assent and their caregivers subsequently provided informed consent. The details of the study have been provided previously by Kiely and colleagues. The MMSE was not used to detect cognitive impairment, but was used as part of the structured mental status assessment with the Delirium Symptom Interview (DSI), Memorial Delirium Assessment Scale (MDAS), and CAM to identify delirium.

In response to this concern, we have revised the text as follow in the Sample section of the Methods:

(p. 10) Of these 611 (13%) displayed CAM delirium. Patients were only included if they were able to provide assent and their caregivers subsequently provided informed consent. The details of the study have been provided previously by Kiely and colleagues. The Mini-Mental State Examination (MMSE) was used as part of the structured mental status assessment with Delirium Symptom Interview (DSI), Memorial Delirium Assessment Scale (MDAS), and CAM to identify delirium.

Minor comments:

P.5 l.80 Please refer to both current “Gold Standards” not only DSM 4!
P.5, l. 83 If I am not mistaken DSM 4 as well as ICD 10 not only capture but define Delirium criteria? Please explain at this point how ICD-10 delirium positive patients fit into the picture?
P.5, l. 84 Why are DSM 4 criteria challenging to apply? The 4 CAM criteria do not seem to be easier to apply, please explain.

Response: We have added reference to the ICD 10 diagnostic criteria. However, neither DSM-IV nor ICD10 criteria played a role in the research we have presented in this manuscript. We have provided more detail regarding the difficulty in applying the DSM criteria. The utility of the CAM is addressed in the following paragraph. We have revised the paragraph identified in the comments above as follows:

(p. 5) Formal diagnostic criteria for delirium were first codified in 1980 in the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders, Version 3 (DSM-III).11 Different definitions have appeared in subsequent DSM versions.12-14 The first appearance of delirium in the International Classification of Diseases occurred in ICD-10.15 While the DSM clearly captures the key elements of the delirium syndrome, the DSM criteria themselves can be challenging to apply diagnostically, both in clinical practice and in research settings, particularly for patients who are not communicative.16 Additionally, the DSM-IV criteria require knowledge of underlying cause before diagnosis can be made. In clinical practice, usually delirium is first recognized and then a search for the underlying cause proceeds. Wide discrepancies in case identification have been reported when different criteria are used.16-18
The goal of our work is to identify the most efficient set of items to determine the presence or absence of each of the CAM features.

What are the authors trying to explain to the reader with this passage? The CAM is often used? That is surely not a quality indicator... The CAM is more often used than the Gold Standards? Medically, speaking this is a frightening thought... Or do CAM positive patients reflect 100% of the DSM -4 and ICD 10 positive patients?

Response: The reviewer is correct; this passage is intended to reflect that the CAM is often used. This statement is not intended to make any quality judgments about the CAM in relationship to gold standard ratings. We have edited the section to clarify this point.

There are many methods for research and clinical diagnosis of delirium, operationalizing either the International Classification of Diseases (ICD) or DSM criteria.

Are there other Delirium instruments that are easier to apply? CAM –ICU takes 90 sec in average, ICDSC takes in average 2 minutes, How about the MDAS, DRS R98, Nu-DSESC etc. they all seem to take much less time with excellent characteristics. Please explain...

Response: The reviewer is quite correct that many instruments take much less time to apply (on average). Shortening the administration time of a standardized CAM administration is precisely the goal of the current research. Some of these existing tools (e.g., the CAM ICU), which was designed for intubated patients who are unable to speak, has been shown to have poor sensitivity when applied to communicative patients. Nu-Desc is a symptom checklist that is based totally on nursing observations and requires no direct interaction with the patient. DRS-98 and MDAS are primarily severity measurements, not diagnostic tools, and take significant time to administer. We chose to not change our manuscript in response to this point, which is better addressed in a review of measures (e.g., Adamis et al. 2010).

I would assume that the binary delirium definition of the CAM is a limitation to the research question / goal at hand. Wouldn’t a severity instrument such as the MDAS or the DRS-R-98 be more helpful in this matter?

Response: In fact, in this research we did not make diagnostic decisions based on the CAM or any other criteria set (e.g., DSM, ICD). So this is not a limitation for the present study. However, to not be misleading, we did use binary classifications of each CAM feature present or absent in this study. This served as the main criterion against which to gauge the information content of candidate screening items. We have addressed this point and indicated text revisions in response to previous comments.
Reviewer: 2

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.

Response: We thank the reviewer for his appreciation of the application of IRT to the development of a screening tool for delirium. However, we believe this is a highly innovative study for case identification as our approach of selecting items is based on item information at the 50th percentile for those who screen positive on the specific CAM feature. This is the first submission of the paper for peer review, as the prior resubmission was in response to our oversight in including a statement that this study had received institutional review board approval. We hope the Editor will recognize the innovation of this study as stated above. In response to this concern, we have revised the text as follows:

(p. 22) The items were chosen in an iterative fashion that incorporates an interdisciplinary perspective from both clinical and methodological expertise in measurement research. The novel approach used in this study for case identification in delirium allows the interdisciplinary team to select items based on item information at the 50th percentile for those who screen positive on the specific CAM feature.

Specific comments (Minor essential revisions):
1. p.3 line 5. line 74: punctuation problem and incorrect capitalization.

Response: We thank the reviewer for this helpful correction. In response to this concern, we have revised the text as follows:

(p. 5) In the US, delirium affects over 2.3 million hospitalized older adults each year at an estimated total annual cost of $152 billion.
2. p.6 line 105 and 106. I suspect that the two references need to be interchanged.

Response: We thank the reviewer for this comment. We have made the correction. We have revised the text as follows:

(p. 6) Item response theory (IRT) encompasses a set of psychometric tools that—among other things$^{26}$—can help in the selection of optimal test questions to shorten instruments.$^{27-32}$

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.

Response: We have included the reviewer's suggestion in the description of IRT. In response to this concern, we have revised the text as follows:

(p. 6-7) IRT can be considered an extension of classical factor analysis$^{33}$ and is a useful tool in test construction because it provides a framework for expressing characteristics of test-takers and test items on a uniform metric.

4. p.7 line 119. $y_{ij}$ 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.

Response: The reviewer's suggestions are now incorporated into the text. In response to this concern, we have revised the text as follows:

(p. 7) Mathematically, IRT can be expressed as a function of a person's response ($y_{ij}$), person-level trait ($\theta_i$), and item parameters ($a_j, b_j$).

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.

Response: We agree with the reviewer and have included one example for the meaning of theta in a clinical context based on the CAM feature of inattention. In response to this concern, we have revised the text as follows:

(p. 7) The unobserved variable (e.g., latent level for the CAM feature of inattention), $\theta$, is often assumed to be distributed normally with mean zero and unit variance. The difference between a person's latent trait level ($\theta_i$) and the item difficulty (or item location, or symptom severity level, $b_j$) defines the probability that a person will display a symptom (e.g., “Trouble keeping track of what was being said,” for the CAM feature of inattention). $P_j(\theta_i)$ describes the increasing probability of displaying indicator $y_i$ with increasing values of the latent trait.
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).

Response: We agree with the reviewer and have streamlined this presentation and eliminated Figure 1.

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.

Response: We have reviewed Reeve et al and have included this reference in the Methods section for the item bank construction, as it is similar to the procedures we followed in standard psychometric analyses. In response to this concern, we have revised the text as follows:

(p. 9) The overall construction and evaluation of the item bank is similar to the National Institutes of Health Roadmap Initiative Patient-Reported Outcomes Measurement Information System (PROMIS). The description of the PROMIS psychometric analysis for item banking is found in Reeve et al.34

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

Response: We agree with the reviewer and we have included the number of items from the source instruments and reference describing how the indicators were derived. In response to this concern, we have revised the text as follows:

(p. 11) The total number of items from the source instruments that were considered in the clinical consensus was 119. The description of the clinical consensus is reported in detail by Huang and colleagues.35

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.

Response: We realize that this was unclear and have now noted that all indicators were dichotomous, as indicated on line 124 in the equation. The indicator response types have been discussed in a previous publication for each CAM feature by Huang and colleagues. In response to this concern, we have revised the text as follows:

(p. 11) All items were dichotomous, as described in Huang et al.35

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.
Response: The check for multicollinearity is for local dependency among pairs of indicators conducted by fitting a weighted least-square factor analysis modeling using Mplus software (version 5.2, Muthén & Muthén, Los Angeles CA). In response to this concern, we have revised the text as follows:

(p. 13) We performed an empirical multi-collinearity check for local dependency among pairs of indicators by which correlations could not be estimated. We fit a weighted least-squares factor analysis model using Mplus software (version 5.2, Muthén & Muthén, Los Angeles CA) to the indicator sets. Multicollinearity is detected through error messages in the Mplus output indicating sparsely populated cells potentially due to logical dependencies among one or more pairs of indicators.

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

Response: We again thank the reviewer for pointing out this typo and have deleted the “/” that was in the text before assumption. In response to this concern, we have revised the text as follows:

(p. 15) When we rejected the assumption of unidimensionality, the indicator sets were split into m sub-sets to achieve sufficiently unidimensional indicator sets for IRT-based data analyses.

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?

Response: We did not develop composites for each dimension, as this was the first step in the development of a more refined screening instrument for delirium. In response to this concern, we have revised the text as follows:

(p. 21-22) The significance of this work is for the future establishment of validated instrument for delirium screening. Our work represents a first step in development of a more refined delirium screening instrument.

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.

Response: We thank the reviewer for this excellent suggestion and have added in the word “curve” to the end of the sentence. In response to this concern, we have revised the text as follows:

(p. 19) We identify this level of the latent trait as the 50th percentile \((\theta_{50})\) curve.
14. p.19 line 409. Tying it back to the predictive use of the tool (Figure 1) could help readers understand the features here.

Response: We have streamlined the introduction and eliminated Figure 1. Nevertheless, as an attempt to aid in clarity in response to this concern, we have revised the text as follows:

(p. 19) Such items would not be useful for screening purposes, even if the assessed symptoms were pathognomonic of delirium. Our goal is to derive a test information curve tuned for screening purposes. We approach this by choosing the items with the most information at the 50th percentile for our item bank. The two highlighted items provide the most information at the 50th percentile of the latent trait distribution in the feature positive group. This is the area of the latent trait of greatest interest for screening purposes.

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.

Response: We completely agree with the reviewer and have clearly stated the importance of this work in the future development of a validated delirium screening tool, as we do not yet present any statistic for the predictive validity in this study. In response to this concern, we have revised the text as follows:

(p. 21-22) The significance of this work is for the future establishment of validated instrument for delirium screening. Our work represents a first step in development of a more refined delirium screening instrument.
Reviewer's report
Title: Selecting Optimal Screening Items for Delirium: An Application of Item Response Theory
Version: 2 Date: 20 August 2012
Reviewer: Joel Coste

Reviewer's report:
The objective of the paper was to identify a short set of indicators to assess the four features of delirium in the Confusion Assessment Method diagnostic algorithm using modern measurement theory. The paper also aims illustrating modern measurement theory applied to clinical assessment. A pool of 135 indicators from established cognitive tests and delirium interview tools administered to 4598 older adults assessed for delirium has been considered. Modern measurement methods have been used, together with classical test theory methods (EFA/CFA) to assess dimensionality assumptions and identify the best indicators for screening for each delirium feature in the perspective of developing a short assessment for detecting delirium. The paper deals with the important topic of modern measurement input to the process of shortening composite measurement scales (CMS). It could turn out to be an important paper for those who embark on shortening such scales. However, there are some areas for improvement, detailed below.

Major Compulsory Revisions
1. The role of modern psychometrics and IRT in item reduction should be further investigated and discussed. Following the application of such methods, items can be selected or removed not only on the basis of their information content (“discriminatory power”) but also on the basis of the level of fit to their respective dimension, and on the basis of closeness or overlap with other items of the dimension. IRT methods allow for a precise diagnosis of the functioning of every item or even every response category of every item. IRT provides a set of interpretive tools (information curves but also item characteristic curves, response probability curves, differential item functioning, etc.) that can be useful for scale refinement or shortening. These tools encourage an interpretation of the results, which is a key aspect for selecting the items to be retained. Furthermore, most of IRT models benefits of the property called specific objectivity (item and person invariance), which allows independent comparison of instruments across several samples and symmetrically independent comparison of individuals across several instruments. This property is particularly precious for CMS shortening, as two instruments (the original and the short-form CMS) have to be used in several samples (development and validation samples at least).

Response: We thank the reviewer for this suggestion to further investigate and discuss the role of psychometrics and IRT in item reduction. We have revised our section of the introduction (pages 6-8) that introduces the concept of IRT and its role in efficient measurement. We have a unique situation in which we have a binary criterion variable (whether or not CAM features were rated as "present" or not that drive item selection. We do consider information and location by choosing items with high information at specific levels of the "person-level trait."
2. IRT methods and factorial analyses, especially confirmatory factorial analysis (CFA), are usually considered complementary to check the effect of the removal of items on the dimensional structure of the CMS. CFA has an important advantage over exploratory factor analysis in that it allows testing whether the initial model remains intact when items are removed. It is important to note here that if the removal of an item alters the dimensionality of the scale, the item should be kept rather than the dimensions modified, in order to respect the underlying conceptual model, which should always be kept in mind during the shortening process. In practice, however, IRT methods and CFA can rarely be simultaneously applied since IRT methods are appropriate to dichotomous/ordinal items and CFA requires normal distribution of the scale items. If the authors wish to continue presenting both IRT and CFA results, they should clarify which technique they used for which scale and item (indicator) distribution of responses and detail how the hypotheses and assumptions of techniques were checked.

Response: We have a somewhat different view of the relationship between IRT and factor analysis. As pointed out by reviewer 2, IRT as an extension to classical factor analysis. Moreover, IRT and factor analysis are isomorphic when the factor analysis is performed on a matrix of polychoric correlations and only one latent variable is modeled. So in direct response to the reviewer's comment that "If the authors wish to continue presenting both IRT and CFA results, they should clarify which technique they used for which scale and item (indicator) distribution of responses" -- our response is that our unidimensional factor analysis results are item response theory results, and more globally our multidimensional factor analysis results are multidimensional item response theory. We also cannot agree with the statement that "In practice, however, IRT methods and CFA can rarely be simultaneously applied" because the two approaches are (in certain implementations, and as we have implemented them) identical. The ordinal dependent variable approach to factor analysis was described by Birnbaum in Lord and Novick's seminal work on IRT, formalized by Christoffersson and Muthén. In our approach, insofar as unidimensionality is an assumption of IRT, we sought first to assess the extent to which our data satisfied this assumption before moving on to formal IRT analyses.

Minor Essential Revisions
1. Whereas pen and paper is still the dominant mode of administration of composite scales, the increased use of electronic administration using Computer Adaptive Testing (CAT) changes the role of IRT methods in the context of selecting items: the number of items that may contribute to a scale score might even be increased with CAT without any corresponding increase in respondent burden. This development should at least be discussed.

Response: We agree with the reviewer. In response to this concern, we have revised the text as follows:

(p. 22) Ideally, in the near future our approach will be enhanced by computer assisted bedside interviewing with well characterized item banks and adaptive testing algorithms tuned to distinct purposes (e.g., grading delirium severity, screening for probable delirium).
2. Check the references -- some seem incorrect eg ref 24 page 27; and spelling of names, eg Gibbon page 16

Suggested additional references
Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16:5-18.

Response: We thank the reviewer and have made the corrections and inserted the selected references.

Our behalf of our co-authors, we are grateful for the careful reviews and helpful comments, and hope that the editors agree with us that revised manuscript is improved and a better fit to the readers of BMC Medical Research Methodology. If we can be of any further assistance in the review of this manuscript, please do not hesitate to contact us (jones@hsl.harvard.edu, francesyang@hsl.harvard.edu).

Sincerely,

Richard N. Jones, ScD
Frances M Yang, PhD

P.S. References Cited in This Note
P.S. References Cited in This Note


