Author's response to reviews

Title: Item response theory analysis of cognitive tests in people with dementia: a systematic review

Authors:

    Sarah McGrory (S.McGrory@sms.ed.ac.uk)
    Jason M Doherty (j.m.doherty@sms.ed.ac.uk)
    Elizabeth J Austin (Elizabeth.Austin@ed.ac.uk)
    John M Starr (jstarr@staffmail.ed.ac.uk)
    Susan D Shenkin (susan.shenkin@ed.ac.uk)

Version: 2 Date: 30 September 2013

Author's response to reviews: see over
Dear Dr Lon Schneider,

Thank you for asking us to submit a revised manuscript to BMC Psychiatry. We are very grateful for the comments of the two reviewers, and believe that by addressing them, the manuscript is much improved.

We have re-run the search to ensure any more recent papers would be included. Therefore the numbers of articles retrieved have increased. No new eligible studies were identified.

We attach a copy of the manuscript with tracked changes, and one without, and provide a point by point response to the reviewer comments below. We hope that the article is now suitable for publication in BMC Psychiatry, and look forward to hearing from you.

Reviewer 1:
This is an interesting study about the applicability of Item Response Theory (IRT) methods in order to improve detection of people with dementia disorders through cognitive tests. .... However, as present knowledge in this particular field is limited to the four studies, this review should be taken as a preliminary approximation to the subject.

We thank the reviewer for their positive comments.

1) It would be helpful if some limitations of the approach were discussed. One example is the implicit assumption that one test taps a single latent construct. What about in those tests or test batteries in which multiple domains/factors are assayed?

We agree with the reviewer and have included a more comprehensive review of the limitations (Discussion: Some limitations of this review should be acknowledged).

2) Similarly no IRT measure has been cross validated and directly compared to the clinical measures. Thus it is unknown if this approach is truly advantageous.

We thank the reviewer for this suggestion and have included this (Discussion: Linearity and the assessment of change in severity, paragraph two).
“This new measure, along with any new IRT measure, would need to be cross-validated and directly compared to existing clinical instruments to ensure this test development technique is truly beneficial.”

3) Can the authors comment on older Chapman and Chapman arguments about how to ensure equivalent difficulty level and true score variance across tests in order to identify differential deficits that are relevant here?

We agree with the reviewer and have included a comment on the historical context of examining item difficulty and discrimination using classical test theory (Discussion, paragraph nine).

“While these studies demonstrate the use of IRT to examine item difficulty and discrimination, the investigation of item differences has also been addressed using classical test theory (CTT). Chapman and Chapman [47] identified the need to study these item parameters in their analyses of specific and differential deficits in psychopathology research, for example, specific deficits in schizophrenia or the analysis of domains or abilities which remain relatively intact in dementia. Chapman and Chapman’s analyses of differential deficits is rooted in classical test theory (CCT) and IRT, as a newer statistical model, offers alternative means of exploring the differential deficit problem. When examining differential deficits between different groups IRT, unlike CCT, can offer estimates of measurement error for different levels of cognitive ability, without having to conduct separate studies, and can establish whether different items or measures are equally difficult.”

4) Some comment on industry’s interest in the approach should be made.

We thank the reviewer for this comment and have included a comment on the industry’s interest in IRT methods (Background, paragraph 14).

“In addition, IRT methodology will be useful to industry in the design of psychometric tests”

Reviewer 2:

1) In my opinion, the authors have adopted too strict inclusion and exclusion criteria.

This review does have strict inclusion and exclusion criteria. The decision was made to uphold these criteria, which we had carefully considered, as it is important to identify all published articles that meet the criteria. Specifically (i) English language – discussed below in point c) (ii) IRT methods – the focus of the article (iii) global cognitive function – discussed in point a) below; (iv) dementia – the results of IRT are likely to be different in people with and without dementia as discussed in the introduction.

Most articles were excluded for more than one reason so to relax one exclusion criteria, e.g. the use of non-English language measures, would not be sufficient to include these articles.
We agree that several of the studies that did not meet inclusion criteria include important and relevant data (e.g. Benge et al. 2011, Balsis et al. 2012, Wouters et al. 2008, Korner et al. 2012, Ideno et al. 2012). We have therefore summarised these results and how they contribute to the findings of the included studies in the discussion (Linearity and the assessment of change in severity, paragraph one (Balsis et al., 2012, Wouters et al. 2008), Information: paragraph two (Ideno et al. 2012), paragraph three (Korner et al. 2012), paragraph four (Benge et al. 2011)).

a. Several important tests of global cognitive function are missing.

The review only included three tests of global cognitive function, but this was a consequence of the inclusion and exclusion criteria, i.e. the tests that have been included in published papers. One study which analysed the Baylor Profound Mental State Examination was mentioned in the discussion but not included because a Dutch version was administered (Korner et al. 2012). We have clarified this (Discussion: Some limitations of this review should be acknowledged, paragraph three).

“This review was limited to analyses of only three global cognitive function; MMSE, BIMCT and ADAS-cog. This was a consequence of the articles meeting inclusion criteria. However, an analysis of the Baylor Profound Mental State Examination, while not reviewed due to use of a Dutch version, was mentioned in the discussion [51].”

b. The authors should also reconsider their choice to only include global cognitive function scales.

We carefully considered the choice to include only global cognitive function scales. We chose these scales to maximise the clinical relevance of the review, as global cognitive function scales are the most commonly used for testing in routine practice. We do, however, agree that domain specific measures are of interest, and the one study identified which applied IRT methods and otherwise met inclusion criteria was discussed [53] (Discussion, Information: paragraph four).

“While global cognitive instruments such as the MMSE are probably the most commonly used measure of cognitive functioning, domain specific neuropsychological tests have been demonstrated to show increased sensitivity to early stages of cognitive impairment than measures of global cognition [32]. Of the seven studies applying IRT methods to domain specific measures identified [53; 54; 55; 56; 57; 58; 59] only one; Benge et al., [53] otherwise met inclusion criteria. This study’s findings are briefly discussed here.”
c. Furthermore, although excluding non-English versions of the instruments is understandable on first sight, it is actually inappropriate after brief reflection.

We thank the reviewer for this suggestion, and we did consider in detail whether or not to include non-English versions of the instruments. However, we did feel this was justified because test language can potentially lead to inaccurate estimations of cognitive impairment and comparing non-English language tests with English language is difficult as the semantic range of items cannot be assumed in translation. In addition, we examined the articles excluded for this reason, and 20 of 21 would not be included even if this criterion was removed (17 for use of non-dementia samples, three written in foreign languages). We have included more detail on this (Methods: Exclusion criteria, paragraph four; Discussion, Information: paragraph three):

“Non-English language versions of cognitive measures were excluded. While several measures, most notably the MMSE, have been translated into many languages for use in different countries and cultures there are concerns over the cross-cultural validity. The language in which a test is administered can affect performance leading to a potential overestimation of cognitive impairment in individuals who do not speak English [35; 36; 37]. Also the non-English language versions administered makes comparison with scales in English problematic because the semantic range of items cannot be assumed in translation [38], for example repeating “No ifs, ands, or buts” corresponds to repeating “We put ones’ efforts all together and pull the rope” in the Japanese version of the MMSE [26] and to a tongue-twisting phrase “en un trigal habia tres tigres” (“there were three tigers in a wheat field”) in the Spanish version [39]. To avoid any potential confounding these articles were not included for full review [40].”

“21 studies were excluded for administering non-English measures. However, all except one were excluded for other reasons also (17 for use of a non-dementia sample, three written in foreign languages). The results of the single study [51] which was only excluded due to use of a Dutch version of the Baylor Profound Mental State Examination are discussed.”

2) The introduction is very methodological.

We agree with the reviewer and have rewritten the introduction with current testing problems discussed in more detail. IRT is then described followed by a discussion on how IRT can be used to address some current testing problems. Some examples of how IRT has done so in other fields are also given. Some methodological details have been removed.

3) Description of studies lacks structure.

We are grateful for the reviewer’s comment and have restructured each study’s description to start with a brief description of the aim of the study and the role of IRT, then describing the study, and its main findings.
4) I have my doubts about the following sentence: "The ICC is the building block of IRT" (Introduction paragraph three). I think TCCs and ICCs are graphical tools to inspect measurement properties of tests and items. The building blocks of an IRT model are its parameters and how these are estimated.

We thank the reviewer for identifying this error and have removed this sentence to provide a briefer more accurate description of the Item Characteristic Curve (Background, paragraph eight).

“IRT can provide two useful measures; difficulty and discrimination, both of which are technical properties of the Item Characteristic Curve (ICC). The ICC is a non-linear regression on ability of probability of a correct response to each item.”

5) The MMSE item label "Pentagram" is incorrect.

We thank the reviewer for identifying this error. The item was correctly relabelled “Intersecting pentagons”.