Reviewer's report

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

Version: 1 Date: 26 August 2013

Reviewer: hans wouters

Reviewer's report:

This is an interesting manuscript that addresses a highly relevant topic for the field of dementia research. Manuscripts such as these are highly welcomed. I really do appreciate the work by these authors. Unfortunately, in my honest opinion the manuscript is not suitable for publication in its current form. I do believe, however, that with the dedicated time and commitment, a completely revised paper might be suitable for publication. At present, I have several major concerns:

# Major Compulsory Revisions

1. In my opinion, the authors have adopted too strict inclusion and exclusion criteria. Of the 340 available publications, only 4 were included, which is approximately ~1%. Although avoiding study heterogeneity justifies adoption of such strict exclusion criteria to allow for a quantitative pooled data-analysis in order to attain meaningful overall conclusions, the authors did not actually pool the data. Since they restrict their ‘systematic review’ to a mere narrative review, the issues of study heterogeneity and data pooling are less important for these authors. The exclusion of so many articles is therefore not justified. At least ~5-10 additional articles could and should be discussed. This is all the more important because at present the exclusion of so many study reports by the authors comes with three significant problems:

a. Several important tests of global cognitive function are missing. The authors restrict their systematic review to the MMSE, the BIMCT and the ADAS-cog. In my opinion, at least the CAMCOG should be discussed. It has been used extensively in the UK and Europe. An IRT analysis of the CAMCOG is actually available and was published by Lindeboom et al. in 2004. If IRT analyses have been conducted on MOCA and 7ms data, then the authors should consider these instruments as well.

b. The authors should also reconsider their choice to only include global cognitive function scales. In the first paragraph, the authors argue: "(...) Therefore the psychometric measures used in dementia need to provide sensitive measurement across the full continuum of cognitive ability. This is essential for early recognition and intervention and to monitor the course of disease progression and effect of treatments". However, a key limitation of global
cognitive function instruments like the ADAS-cog and the MMSE are the ceiling effects which come with a decreased sensitivity to early cognitive decline (see Harrison et al., 2007). Actually, several studies including the study by Harrison et al. show that domain-specific neuropsychological tests are more sensitive to early impairment than global cognitive measures. Therefore, rather than excluding neuropsychological tests, it would be better to distinguish between global cognitive function measures and neuropsychological tests and to include the latter.

c. Furthermore, although excluding non-English versions of the instruments is understandable on first sight, it is actually inappropriate after brief reflection. Many non-English observation studies of cognitive function were done, which constitute a major contribution to the progression of dementia research. Examples of important but non-English work in the field are the work by Bäckman et al., the Rotterdam Study by Breteler, the AMSTEL Study by Jonker and Schmand and the REAL.FR Group of Vellas. Also, a multinational RCT using the ADAS-cog was conducted by Wilcock et al. Therefore, it would be better to discuss non-English language as a potential problem rather than adopting it as an exclusion criterion.

2. The introduction is very methodological. Many properties of IRT analyses are described. However, it would make more sense to first discuss current testing problems in more detail, then discuss IRT in more detail and then address the question(s) how IRT models will resolve these current testing problems.

3. Description of studies lacks structure. What is the key message that the reader should bear in mind after reading each paragraph? It would be better to organize synthesis of study findings around themes. The manuscript would become particularly clear if the chosen themes would correspond to current testing problems as described in the Introduction.

# Minor Essential Revisions

1. I have my doubts about the following sentence: "The ICC is the building block of IRT" (Introduction paragraph 3). 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.

2. The MMSE item label "Pentagram" is incorrect. I think the authors have the constructional praxis item in mind that asks patients to copy two overlapping PENTAGONS? Verify every item label as many readers have detailed knowledge of most MMSE items and will be confused by incorrect labels.
Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

I declare that I have no competing interests