Reviewer’s report

Title: The predictive value of mental health for long-term sickness absence: The Major Depression Inventory (MDI) and the Mental Health Inventory (MHI-5) compared

Version: 4 Date: 28 June 2013

Reviewer: Niels Smits

Reviewer’s report:

Major Compulsory Revisions

I still see the term ‘prevalence’ is being used in the introduction when citing at risk estimates, which is awkward because in the reply to my earlier comments it was stated: "We accept the critique that the number of persons scoring positive does not constitute a true prevalence since our tests is not gold standard". Although in the papers the authors refer to, it is claimed that they estimate prevalence, it is the estimated level of the test they actually provide statistics on. So, again, these papers are incorrect in their claims. Moreover, the journal BMC Medical Research Methodology being a Methodological journal, these incorrect statements should not be made in the current paper. The following strategy should be used to deal with this. Start the introduction saying (i) policy makers need prevalence estimation etc., (ii) what is available is gold standards such as clinical interviews and self report questionnaires, which are much cheaper and more practical than gold standards; (iii) self report questionnaires are valuable and although only at risk proportions may be estimated (next you can cite the statistics of refs Croatia, Canada and Denmark, do not use the terms prevalence and incidence, but the level of the test equivalents!), they are still very useful for sifting respondents who are at high risk of having low mental health out of the population. (as a possible footnote iv, one may argue that gold standard and screener may be jointly used to do two phase sampling (a cheaper method than administering gold standard to everyone), still giving valid estimates of prevalence).

The last paragraph on p. 3 shows what one can do to validate a mental health instrument. Could you please give a reason why one would give this overview, and how it relates to what was written earlier?

On p. 6 you state: "According to McNamee[5] is a rule of thumb that specificity added with sensitivity should be above 1.6, if the questionnaire is used for screening". This was taken out of the context! In MacNamee this was true within the context of two phase sampling, but not necessarily outside this context. Please find another rule of thumb reference. Perhaps look at the Area Under the Curve (see, e.g., Doi: 10.1016/j.biopsych.2005.09.014).

Minor Essential Revisions
Through the paper the word `continuous' is used, combined with scale. Please delete the word continuous (and note that `scale' is enough). These data, which are sum scores of Likert scale items (which therefore have a scale between ordinal and quantitative) are not continuous. Continuous would mean that all theoretical values (e.g., 3.558954) would be possible. Please read: https://en.wikipedia.org/wiki/Level_of_measurement

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests