Reviewer's report

Title: Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data

Version: 1 Date: 11 October 2010

Reviewer: Mike Kenward

Reviewer's report:

Major Compulsory Revisions

1. All missing data problems have structure, and there is some theory to guide us. This could be emphasized more in this paper. In particular, it is useful to make the distinction between so-called principled and unprincipled methods of handling missing data. Simple imputation methods can be described as unprincipled if they rest on assumptions (often implicit) that contradict the analysis model, irrespective of the missing value mechanism. The analysis model here is the Rasch model, so obvious examples of unprincipled imputations in the current setting are the worst case and mean score methods.

I suspect that most, but not all, of the others are similarly unprincipled. In general, unprincipled methods will lead to bias, although in some settings this will be small. Thus we know that some methods will be biased, even under the MCAR mechanism and this should be made clear in the methods section.

Similarly we know that principled methods will be consistent, at least under MCAR. The likelihood methods of NOIMP and LD are examples. We know these will be consistent under MCAR; what we do not know is how much bias will appear under the non random missing value mechanism.

2. A major issue with single imputation methods is that unless very special provision is made in subsequent variance estimation, they will lead to an overestimation of precision:

data are being "made-up" and this fact is subsequently ignored.

This is mentioned in passing in the discussion but must be made more central. Unless estimation is required without subsequent inference, and this is rare, all the single imputation methods fall down at this hurdle. Simply, they should not be used. And we don't need to use them. We can use direct likelihood methods, or if this is not convenient, we can use multiple imputation. There is now a very good literature on this, and I was not sure why this was not discussed further, although I would not expect it top be used here. The reason is that without explicitly introducing a non-random missing-data mechanism it will lead to analyses that are very similar indeed to the likelihood based analysis NOIMP, in fact they will be asymptotically equivalent as the number of imputations increase. Where there
is no obvious likelihood route, for example when estimation is ad hoc with incomplete data (e.g. Loevinger’s H) then use multiple imputation and estimate the target quantity from the completed data sets. The reader needs to be made aware of these possibilities and links.

3. We therefore know what will work under MCAR, and what will be potentially biased. The interest therefore is how these methods will perform under the non-random missing-data mechanism. What struck me most strongly about the results was how robust the methods were to variation in the correlation that governs the non-randomness of the missing-data mechanism. Only in WORST was there non-trivial dependence on this parameter. The important question is why? Is there something about the chosen mechanism which implies so little impact? Or should different parameter values be used in model (3). Until we have answered these questions we don’t know whether we have been testing these methods under a serious non-random mechanism. A model needs to be used that has some non-trivial impact on the behaviour of the likelihood analysis at least, because this represents a gold standard under the MCAR mechanism. This means that it is difficult to draw any comparative conclusions from the results given about the behaviour of these methods under MNAR. In the other hand, the correct conclusions under MCAR come from theory, and don’t need simulation: i.e. where possible use maximum likelihood (NOIMP), don’t throw away data (LD) and don’t use single imputation (all the others). If methods need complete data for calculation, then use multiple imputation.

4. Because of these points I think that the title needs changing accordingly.

Minor Essential revisions

5. Methods, last paragraph:

What happened in the simulation to datasets with more than 50% missing - were they discarded?

6. Simulation design, paragraph 3.

"testing ignorability" I don’t know the paper by Holman and Glas, but at the least this sentence needs to be phrased very carefully. We cannot establish from the data under analysis that the missing-data mechanism is ignorable. We can only show that the probability of being missing depends on observed data, which contradicts MCAR. An absence of such an association does not imply MCAR, or rule out MNAR; and the presence of such an association does not necessarily imply MAR.

Discretionary revisions

7. To maximise the context of the work, it would be useful to point out that the Rasch model defined in (1) is just a particular and simple example of a generalized linear mixed
model. It is worth checking to see if there is literature on missing data for the logistic version of a GLMM, outside the psychometric field, although I am not aware of any that replicates the authors work, and of course the specific target quantities such as Levinger's H and PSI are specific to this setting.

8. As well as bias it would be of interest to see how precision is affected by the chosen methods. As there is a potential for bias, this could be assessed through mean square error.

Minor issues Not for publication

There are innumerable grammatical problems and typos - too many for me to list. The paper needs to be read carefully by a native English speaker.

**Level of interest:** An article of limited interest

**Quality of written English:** Not suitable for publication unless extensively edited

**Declaration of competing interests:**

I declare that I have no competing interests.