Reviewer's report

Title: Dealing with Missing Data in a Multi-question Depression Scale: A Comparison of Imputation Methods

Version: 1 Date: 20 September 2006

Reviewer: Paula Diehr

Reviewer's report:

General
The revised version of the paper is more complete and easier to read. This is an important and complex problem, and it is difficult to derive a simulation that tests all of its possible facets. The authors have made a good start. However, there are still problems in three areas: the missingness mechanism, the attribution of the multiple imputation results, and the “impression” of the findings.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Missingness Mechanism:
The three usual missingness mechanisms are Missing completely at random (MCAR), Missing at random (MAR) and Missing not-at-random (MNAR). Most simulation comparisons of imputation methods examine performance of imputation methods in all three cases. One might argue about the operational definitions, but a plausible one would be that the MCAR is the one they have used, that MAR would set the probability of being missing to be a function of something that was known (for example, the probability that item 6 is missing depends on the response to item 1 which is known), and MNAR would be a function of the value of the missing item itself (for example, item 6 would be missing half the time if the true answer was yes but only 5% of the time if the true answer was no). Especially since this journal is focused on methodology, this paper needs to describe these mechanisms and develop a missing data set that has each type of missingness mechanism, to provide a better evaluation of each method. The case they now emphasize, MCAR, seems trivial.

Multiple Imputation:
The authors state that multiple imputation performs the best. However, as noted in the previous review, what they have actually shown is that the average imputed value coming from a particular “experimental” SAS program performs well. That program is not described in great detail, but presumably has a procedure for estimating the missing value (which I call the regression part) and performs it 5 times (which is the “multiple” part). It may be a misnomer to attribute the good performance of the estimate to its being “multiple”, when only the “regression” part is used in their simulation. The “multiple” part is relevant only because the mean of 5 estimates may be more stable than a single estimate.

In response to a similar comment in the earlier review, the authors have added a “single regression” imputation, which does outperform some of the simpler methods. This regression was described in a single sentence, but was presumably quite complicated to program, since every person would have different missingness patterns and thus different covariates available. This should be explained. (Whether previously imputed missing observations are used to impute a new missing observation is also not explained). At any rate, although the estimate performed well, there is no reason to believe that this regression estimate is the same as would result from a single imputation from the SAS program, and likely it is not.

The question of whether the SAS program performed better because of its regression part or its multiple part is left undetermined. The authors could investigate this problem in two ways. One would be to run the SAS program with only a single imputation (that is, set the parameter to 1 replication instead of 5). If this estimate was also superior to other estimates, then it would be clear that it was the “regression” part that outperformed other methods. (An alternative and less desirable approach would be to add a random observation drawn from the residual distribution to their “single regression” estimate, do this 5 times, and compare the mean of those estimates to the SAS estimate.) This would help to separate the better performance of the regression part from the benefits of multiplicity. The expected value would be the same with or without the multiplicity, but there might be some advantage in taking the mean of 5 estimates. If it appeared that there was such an advantage, it would be interesting to see whether 2 or 3 or 4 or 10 imputations provided different advantages.
Over-all findings:

Finally, the authors may be making too much of the small differences in the kappa statistics. It would require creating a great many datasets, calculating a kappa for each, and then calculating a paired t-test of the kappa statistics to permit saying whether one kappa was significantly higher than another, which is probably beyond the intended scope of this paper.

The authors summarize their results in the abstract as showing that multiple imputation is best. However, the excellent performance of the “individual mean”, which always had the first or second highest kappa and lowest missingness, was even more striking. The individual mean is so much simpler than using either the single regression or the SAS program that this finding (with caveats mentioned in the discussion section) seems to be the most important one. This finding (if it holds true for the proposed new MNAR missingness dataset) should be given even more emphasis, as it may help inexperienced researchers. This result is appropriately stated in the abstract, but the “impression” that is given is that the MI method should be chosen, when its superiority is small and possibly not statistically significant.

------------------------------------------------------------------------------------------------------------------
Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

------------------------------------------------------------------------------------------------------------------
Discretionary Revisions (which the author can choose to ignore)

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

'I declare that I have no competing interests'