Reviewer's report

Title: Should researchers use single indicators, best indicators, or multiple indicators in structural equation models?

Version: 2 Date: 10 July 2012

Reviewer: Christian R Geiser

Reviewer's report:

I have read the authors’ responses to my comments and the new version of their manuscript carefully and with great interest. I appreciate the authors’ detailed explanations in their response letter. A number of things have now become clearer to me with regard to the authors’ way of thinking. Also, the manuscript has improved, particularly the introduction, which is now more (albeit still not fully) understandable to readers not familiar with the related SEMNET debates.

Compared to current practices, the paper offers a radically different perspective on selecting and specifying indicators in structural equation models (SEMs) with latent variables. It calls for a more rigorous theory-driven approach to setting up SEMs with which I strongly concur. The paper makes a number of very important points and should be useful to advance the field in several ways.

One of my remaining concerns is that, given the radical differences between the approach advocated by the authors and the approach currently pursued by the “mainstream” (including both methodologists and substantive researchers), the current write-up is still too much targeted towards the specific audience that follows (or is at least somewhat familiar with) the related SEMNET debates. As an example, the description of the close vs. exact fit issue on p. 3 is cryptical and definitely hard to follow for somebody not involved in the corresponding SEMNET exchanges. Is this not a side issue for the current indicator discussion? Is it at all relevant here? Could it not be dropped?

Also, I found that although the authors explained many things very well in their reply to my initial comments, most of these very useful explanations did not find their way into the revised manuscript. The authors now mention “misconceptions” briefly at the end of the manuscript (p. 33) and it sounds a little bit like those were just my personal “misconceptions”. But their explanations come at an odd place and are not at all integrated into the general discussion. It is awkward to say, “OK, were done with the main discussion now and here are some of Christian Geiser’s misconceptions.” Moreover, the explanations given on p. 33 are way too sparse for the reader to fully understand the issues. I am fairly certain that the “misconceptions” that I raised in my previous review are not just my personal ones. In fact, many other researchers in the field share these “misconceptions”. So I think that it would be necessary and worthwhile to address these points in the manuscript in the same helpful way in which the authors did this in their response letter. This would help readers to better understand why the authors’
approach is seen as superior over currently established procedures.

Regarding the "misconception" of breadth of latent variables, my point was that many researchers at least in psychology would not accept a latent variable labeled, say "depression" that was operationalized with just a single item, say "I feel sad all the time". First, single items usually do contain more random measurement error than do aggregates based on multiple items. Of course, the meaningful use of an aggregate requires the assumption that the items are at least congeneric (and an appropriate weighting scheme if the items are not at least essentially tau-equivalent), but with more than 3 items, this assumption can be tested. Many people would think that if we can find a set of congeneric items, it would generally be better to use the set of items rather than a single item, because the aggregate would be less contaminated by random measurement error.

Second (and probably more importantly), one basic idea of factor analytic measurement models (I know the authors are not a big fan of factor analysis, and I'm not saying that I am one) is that random measurement error AND specific factors can be separated from common factor variance by using MULTIPLE indicators of a more broadly defined construct, say depression. Call it breadth or whatever else, but to me it seems that this is seen as one of the major strengths of the SEM approach and of using MULTIPLE indicators by many people in the field. Again, I'm not saying that I share this view. However, this point is likely of interest to the readers, and a main point of controversy. Unfortunately, it is not sufficiently addressed in the current version of the manuscript.

Additional, more specific comments:

- The authors largely ignored my previous comments on the reliability-validity discussion. Even though they make it clear in their response letter that “the issue is validity, not mere reliability”, this still does NOT become clear in the paper. The authors still use expressions like “true-score-like latent variables”, “true score”, and “true-score variance” (e.g., on p. 3, p. 12, and p. 14) and “measurement error” (in various places in the manuscript, e.g., on p. 14 and p. 25), which to many readers will suggest that the issue IS reliability as described in classical test theory (CTT), from which these terms are borrowed and which focuses on reliability, not validity. As another example, on p. 13, the authors state that “attaining a valid model specification might require specifying y5 as less reliable.” (emphasis in the original) As I had tried to make clear in my first review, this will likely be confusing for readers, because typically, a correction for attenuation IS based on CTT-based measures of reliability, and NOT validity estimates. So what the authors propose is something completely new and unusual, and hence terms should be used in a very clear and consistent manner. Furthermore, as an issue raised previously, Borsboom and Mellenbergh cautioned against mixing up issues of reliability and validity on both the conceptual and empirical level, and this issue is still not discussed by the authors.

- I insist that the way the authors currently discuss their “real examples”(starting on p. 28) is rather unhelpful for researchers who want to understand how to
derive reasonable estimates for the fixed error variances in practical applications. Especially the current description of the loop simplex model (Example 1) is completely useless, as the authors do not explain anything related to the fixed indicator residual variance issue for this example, but instead only refer to a book chapter. I can understand the authors’ statement that every research situation is unique and that general guidelines are therefore hard to give. But the authors must have gotten their ideas for how to fix the residual variances in their examples from somewhere and even describing that would be helpful, because it gives researchers an idea of how experts proceed in these situations, even if this cannot be directly transferred to their own research problem. Otherwise, the “real examples” will just fill space, but will be of little or no utility to interested readers.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

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.