Reviewer's report

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

Version: 1 Date: 23 March 2012

Reviewer: Christian R Geiser

Reviewer's report:

In this article, Leslie Hayduk and Levente Littvay are concerned with the question of whether researchers using structural equation models (SEMs) with latent variables should use a single best or multiple indicators (observed variables) per latent variable. Hayduk and Littvay argue that use of a single best indicator is preferred, because such a strategy allows for the inclusion of more latent variables in a single model compared to models that employ multiple, potentially redundant indicators per latent variable. The authors recommend routine use of single-indicator models in which the path from the latent variable to the indicator is fixed to 1.0 and the indicator residual variance is fixed to a value chosen based on theory, prior knowledge, and/or best guesses about the (un)reliability and/or (in)validity of the indicators.

The topic of this article is highly relevant to all researchers using SEMs with latent variables, as use of fewer indicators can save a lot of time and money in practice. An important contribution of this work is that it stimulates our thinking about the causal connection between latent variables and their indicators (and about causal connections among latent variables) as well as the question of how SEMs can be specified in a more theory-driven and holistic way (in which measurement issues are not separated from the “structural” part of the model). To date, many researchers follow “rules of thumb” (e.g., “at least 3 indicators per factor”) without clear theoretical or substantive reasons for doing so.

That being said, I found the following aspects of the manuscript to be in need for “major compulsory revisions”:

1. Methodologists have tended to make the opposite recommendation in the past (i.e., “having more good indicators per latent is better”), partly supported by simulation studies (e.g., Marsh et al., 1998). The authors do not at all discuss arguments in favor of using more indicators per latent (e.g., improvement of the identification status, improved estimation quality, more broadly defined latent variables, etc.). In my opinion, a more balanced discussion of this issue would not hurt, but rather improve the presentation. Of course, the authors are free to recommend what is better in their opinion, but the readers should at least be informed of potential limitations/downsides of the single-indicator approach.

2. For example, one limitation of the single-indicator, fixed error variance approach advocated by the authors is that researchers may find it difficult in practice to choose appropriate values for fixing the indicator residual variances.
How do we know in practice how much of the variance in our indicator reflects variability caused by the latent variable in question versus other latent variables and random measurement error? This will often be very difficult to figure out in practice, even if researchers work with already established items/scales and/or have a relatively strong theory about their latent variables and the corresponding measures. The authors refer to other publications in their article for a more detailed discussion of this problem; however, given that the problem of how to properly specify the indicator residual variance is at the heart of the authors’ approach, it would be preferable if the authors could say something more concrete about this issue in the paper. Ideally, such a discussion should be based on a concrete example. The article is in large parts very theoretical anyway; a running empirical example would help a lot in making the paper easier to read and the procedure more concrete and useful for both methodological and applied researchers. The authors have included two substantive examples at the end of their paper. However, I found the discussion of these examples in the text to be very sparse and not very helpful to better understand how we should find appropriate values for fixing single-indicator residual variances.

3. Related to the previous point, in my opinion, the authors do not discuss potential problems related to a misspecification of the indicator residual variance in enough detail. Clearly, an over- or underestimation of the indicator residual variance can lead to greatly distorted results in the structural (latent variable) model. The authors recommend sensitivity analyses to examine the impact of choosing different values for the indicator residual variance on parameter estimates in the structural part of the model. However, in practice, researchers will often have no way of knowing whether the resulting estimates of the structural model are valid or not. This is especially true in the “underidentified-if-freed” case, in which a model would no longer be identified if the error variance was freely estimated. Moreover, different path coefficient estimates in the structural model (resulting from different choices of the indicator residual variance) may all be “proper” and “plausible”. How does the researcher decide for one or the other? This problem will be even aggravated when researchers test different conflicting theories about the signs and/or magnitudes of the effects in the structural model. Different choices of the indicator residual variance may favor one or the other theoretical approach. How would a researcher decide between the competing theories when the answer depends on his choice of the indicator residual variance(s)? The authors should discuss this problem in more detail in their article.

4. The authors recommend the use of a single indicator per latent variable, even if this indicator represents a single item. I found this recommendation to be somewhat problematic, given that single items often have very low reliabilities to begin with. As a consequence, much of the variance in such a measure would simply reflect random error variance. If additional (“reliable but construct-unrelated”) variance has to be separated out as part of the error variable, then not much may remain. Furthermore, many researchers at least in psychology would argue that a single item is not sufficient to fully capture complex constructs such as, for example, depression, extraversion, or intelligence. My point here is that the recommendation of using single-item
indicators may often be problematic in practice. Of course, the authors may argue that the “single best” indicator should be chosen (that does have sufficient reliability and validity), but this may still be difficult to achieve with a single item.

5. Borsboom and Mellenbergh (2002)—in response to an article by Schmidt and Hunter (1999)—forcefully argued against the routine use of a correction for attenuation that takes into account both unreliability and invalidity of the measures at the same time. One of Borsboom and Mellenbergh’s main arguments against such a simultaneous correction was that true scores as defined in classical test theory can and should not be equated with construct scores. Furthermore, Borsboom and Mellenbergh argued that issues of (un)reliability and issues of (in)validity should be kept separate, because, according to these authors, even though reliability is quantifiable and hence mathematical corrections for unreliability are possible and may be useful, validity is more of a qualitative concept (according to Borsboom & Mellenbergh, a test is either valid or it is not, but there are no “degrees” of validity) and mathematical corrections for invalidity may not be possible. It looks like the authors of the present paper take on a different position that is more in line with Schmidt and Hunter’s approach, even though Hayduk and Littvay’s approach is model-based and may imply testable consequences for the observed data. My point here is that the authors of the current paper should make it clearer which theoretical position they take in the contemporary reliability-validity discussion (see also Borsboom, Mellenbergh, & van Heerden, 2004). Furthermore, they seem to subsume both unreliability (random measurement error) and systematic error under the term “measurement error”, which is at odds with Borsboom and Mellenbergh’s view, according to which we should not equate true scores and construct scores and not mix up issues of reliability with issues of validity. Again, the reader should be informed about these alternative views, and the authors should make clear where they stand in relation to other researchers who have dealt with similar issues.

References

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30, 505-514.


Level of interest: An article of importance in its field
Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests.