Author's response to reviews

Title: Shame for Disrespecting Evidence: The Personal Consequences of Insufficient Respect for Structural Equation Model Testing

Authors:

Leslie A. Hayduk (lhayduk@ualberta.ca)

Version: 3 Date: 19 September 2014

Author's response to reviews: see over
Reviewers' report

Title: Shame for Disrespecting Evidence: The Personal Consequences of Insufficient Respect for Structural Equation Model Testing

Version: 2
Date: 16 July 2014
Reviewer: Peter Molenaar

Reviewers' report:

This manuscript presents arguments to use the likelihood ratio test to assess the goodness of fit of structural equation models, recommends not to use model fit indices to assess goodness of fit, and discusses the possibility that the amount of (presumably significant) misfit does not scale monotonically with actual model misspecification.

_The article concerns striving to assess the “goodness of the model” and NOT just assessing the “goodness of fit”. The difference is important because we seek properly structured models, and not just models that fit. It is possible for causally-wrong models to fit, and hence assessing the “goodness of the model” must be more stringent than assessing “goodness of fit”._

I have the following remarks (in arbitrary order).

1. The concept of “model” is central in the manuscript. Yet it is not defined in a logically appropriate way. Defining a model as a set of linear equations making up a standard structural equation model is not sufficient (see for instance Mary Hesse, 1966, Models and analogies in science). Also the all-important relation between model and reality is not addressed at all. Since Nancy Cartwright’s 1983 book entitled “How the laws of physics lie” it is clear that the relation between models and reality has to be addressed with special care. I have the impression that the lack of clarity about the model concept in this manuscript also gives rise to what I conceive as a grave misinterpretation of G.E.P Box’s statement referred to on page 10, lines 20-21.

_The model hypothesis underlying structural equation model testing has now been described as striving to match the processes producing the observed_
variables by adjusting the model until its features are consistent with the available evidence. This is not really a definition, but it does situate this article within Cartwright’s work as seeking “the reality behind the appearances” (Cartwright, 1983:1). I have not incorporated a reference to Hesse or Cartwright because I could not see any arguments which depended upon the Hesse’s unresolved tension between mandatory-analogy and absence-of-analogy, or Cartwright’s concern for the fundamental nature that resides behind appearances. The issue of respecting evidence of inconsistency between a model and relevant data need to be viewed as an issue of analogy or fundamental nature – it can be viewed as an issue of evidence. Where that evidence may ultimate lead, or how it will lead us there (Hesse and Cartwright style concerns), stand as separate concerns from what constitutes evidence of undetermined problems in structural equation models.

To ensure that Box has been properly referred to, four new references to Box have been incorporated, two of which relate to a structured world residing behind observed measurements.

2. On page 6, top, it is stated that “the amount of ill fit between the data covariance matrix and the model-implied covariance matrix sigma does NOT confidently correspond to the seriousness of the problems in a model’s specification”. What exactly is meant by “confidently”?

The term confidently has been replaced with more appropriate descriptors.

The examples given in the text following this statement are nor helpful to assess this.

The text that followed provided arguments, and not “examples”. Multiple examples of incorrectly specified models that fit perfectly are given in Hayduk (2014) and are referred to, but these examples are not repeated here. The reviewer’s concerns about specific arguments are addressed below.

The example involving the Ledermann bound is not helpful because in my understanding this bound is an upper bound on the number of factors that can be identified in an exploratory
orthogonal factor model. The Ledermann bound for 4 observed variables is 1; fitting a 1-factor model to the implied (4,4)-dimensional covariance matrix yields a likelihood ratio test that asymptotically is chi square distributed with 2 degrees of freedom. That likelihood ratio almost always is not zero, contrary to what is suggested on page 6, line 6.

The reviewer is technically correct because the Ledermann bound does not guarantee a perfectly fitting model unless the Ledermann bound number of factors is supplemented with a saturating number of error covariances. The omission of the supplemental covariances has now been corrected. Thanks, this was a helpful clarification. In the four-indicator one-factor case, no further factors can be estimated (as indicated by Ledermann), but adding two measurement error covariances to the Ledermann-bound factor model results in perfect fit – even if the true/worldly model is extremely unlike a factor model. Thanks also for correcting the spelling of Ledermann.

The next example concerns model equivalences, in particular the observation that reversing the causal direction of paths in, for instance, a quasi-simplex model, yields exactly the same model fit (same values for the likelihood ratio test and model fit indices). This kind of model equivalence can be detected rather simply, given an initial model.

The fit-equivalence can be detected (as equal fit). But the issue is that the equal-fit precludes fit from determining which of the CAUSALLY DIFFERENT models corresponds to the worldly structure. One model will be wrong, but the lack/absence of ill fit of the wrong causal model signals that even wrong models can fit perfectly. Fit-equivalence is not the same as model-equivalence.

I fail to understand how exactly this corroborates the statement about the implications of the amount of ill fit.

This section has been reworded to highlight the difference between model-equivalence and fit-equivalence.

Similar remarks can be made with respect to the third example involving saturated models have zero degrees of freedom for the likelihood ratio test. On
the top of page 7 the final example is given: “... factor models that provide perfect covariance fit despite being seriously causally misspecified ...”. This is in my opinion the only possibly relevant example.

The revisions noted above should now make this only one of several ways of seeing that perfect fit is not identical to perfect model specification.

And it immediately raises the question: how generic is this example? Also, what more can be said about the simulation models referred to? Can we learn from these simulation models how to detect such cases with real data? But most of all: how can this example help further specify the rather fuzzy term “confidently”?

The diversity of the misrepresented models implies it is unlikely that a single, or even few statistical-indicators will be able to differentiate among the multiple fit-equivalent yet causally misrepresented models, but that is an issue beyond the current article because the current article is arguing for improved attention to models we CAN detect as problematic. Diagnostic investigations are possible for these models, and the relevant diagnostics have now been briefly summarized – largely in response to comments by Reviewer 2.

3. The “amount of ill fit” not scaling monotonically with the degree of model misspecification also is the reason why so-called fit indices are categorically rejected as suitable for the assessment of model fit. Perhaps. But suppose that the example concerning factor models that provide perfect covariance fit despite being seriously causally misspecified, considered in my previous point, is indeed generic enough to take seriously. Then why does this example also not invalidate the chi square goodness of fit test?

The article now clearly reports that while $\chi^2$ is not perfect at detecting specification problems, it nonetheless remains the strongest/best available test. Given that $\chi^2$ is stronger/better than all the indices at detecting problems, its use remains the best practice open to researchers.

4. On page 9, line 4, reference is made to a 1999 paper by Hu & Bentler, as evidence that “No known pair of indices was capable of reliably detecting a
variety of relatively simple coefficient misspecifications built into even a limited set of misspecified factor models”. I’ve reread the Hu & Bentler paper concerned and fail to find the evidence for this strong statement in the manuscript. In their Conclusion and recommendation section, Hu & Bentler state, for instance: “The analyses also suggested that some of our combination rules, based on Hu & Bentler’s (1997) two-index presentation strategy, were able to retain relatively acceptable proportions of simple and complex true-population models and reject reasonable proportions of various types of misspecified models in most conditions”. Quite a different conclusion by Hu & Bentler, it would seem, then suggested in the manuscript.

The sentence was indeed incorrect, and has been revised appropriately.
Thanks for correcting this.

In sum, I am congenial with the general message of this manuscript, namely that structural equation models should be appropriately assessed regarding their fit to the data.

I think model testing constitutes an “appropriate assessment” that is made possible by fit (or lack there of) to the data. This issue is not the fit. The issue is whether or not the fit impugns the model.

But I find the categorical rejection of alternative model fit indices in this context not convincing at all.

The 10 sections of this article address inappropriate challenges to testing structural equation models. Once the objections to SEM testing are eliminated as unfounded, and researchers attend to testing their models, the fit indices will have to re-develop their own justifications, or die for lack of defensible use – which constitutes a paper for another day.

I also think it is essential to present a much more extensive definition of what exactly is a scientific model (like structural equation models) and how one should conceive of the relation between model and reality according to the best insights currently available.
These issues were addressed in my comments above, and the corresponding revisions to the manuscript.

Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests: I have no competing interests

Reviewer 2

Reviewer's report
Title: Shame for Disrespecting Evidence: The Personal Consequences of Insufficient Respect for Structural Equation Model Testing
Version: 2 Date: 8 September 2014
Reviewer: Emily Blood

Reviewer's report:

Review of "Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing.
Overall the article is well written and makes the admirable point that the arguments against the chi-square test of model fit are not persuasive enough to warrant not using the test. While the point is sound overall, some improvements could be made to the discussion.

Improvements are laudable, so we shall have to see what this reviewer proposes as improvements.

The author begins the Background by noting the ‘too quiet’ revolution of Rodgers on the topic of hypothesis testing, so I assume he wants to avoid the same fate for the current discussion, but being too ‘loud’ and, in my opinion, a bit inflammatory and drastic at times (e.g. recommending “public shame”) might instead discourage readers from adequately hearing the message which is overall thoughtful and important.
Whispering of the message was tried and was unsuccessful. A clear and
direct message is now being sent. I think most readers will attend to clear
and direct messaging. Some people will not “hear the message” no matter
how it is presented. I will stand behind clear and direct messaging. I expect
that public shame will indeed be accorded scientists failing to attend to
evidence of problems in their models because disregard of relevant evidence
is not taken lightly by scientists.

Not testing or misinterpreting statistical
hypotheses and ignoring useful statistical techniques is certainly not to be
recommended, but I think a more measured discussion can be had on this topic.
One example of this possibly unnecessary inflammatory tone is the use of the
Sorbom quote to characterize fit indices and the researchers who developed
them as just trying to placate researchers concerned with significant chi-square
tests. A more balanced view is that they were attempting to improve upon some
of the notable imperfections of the chi-square test.

The Sorbom quote directly reports the reason one statistician developed a
specific index – and it was indeed to “placate researchers concerned with
significant chi-square tests”. Other statisticians may have had other
reasons for developing their indices, but the current article is not about the
indices – it is about the $\chi^2$ test, and why the $\chi^2$ test ought not be
disregarded.

The current article addresses specific SUPPOSED imperfections in $\chi^2$
but since none of these are mentioned in these reviewer comments, and no
new/different imperfections are claimed, we will have to leave the various
indices to champion their own improvements/contributions.

Indices become TESTS when anyone specifies an index-value that
(supposedly) differentiates between an acceptable and unacceptable
MODEL. No suggested index value would routinely reject models accepted
by $\chi^2$. But how should a researcher respond if $\chi^2$ reports their model is
significantly inconsistent with the evidence (for unknown and unexamined
reasons), when accompanied by index reports claiming the model is
acceptable? The article is consistent with the view that the stronger
evidence of problems must be respected and diagnostically investigated no
matter what any index reports, because weaker index-testing does not
nullify or overrule the results of stronger testing. The statisticians inventing
the various indices may have had very good intentions, but their intentions will not nullify the requirement to respect stronger evidence of problems.

Many of the arguments included here come to one point: proper presentation and interpretation of results be it the model fit tests, other fit indices, characterizations of the null hypothesis in the Chi-square test, or the consequences of statistical assumptions.

Proper presentation and interpretation is required but the article goes beyond that by indicating what constitutes improper behavior – namely failure to report and respect relevant evidence – the $\chi^2$ test results.

This is not a point that is specific to SEM, but is point more generally applied to statistics as a whole that results from problematic misunderstanding and presentation of statistics done by practitioners rather than fundamental problems with the statistics themselves.

On this point I will disagree with the reviewer. In this instance, statisticians are largely at fault for purveying problematic and misleading recommendations– if one counts people like Michael Browne, and Karl Joreskog and Dag Sorbom as statisticians. For example, see the discussion of Browne, MacCallum, Kim, Andersen and Glaser (2002) cited in the article, and consider the Sorbom quote as a direct report of what a statistician/Joreskog was doing.

This point could be made more clearly, and more helpful to readers would be to lay out (or at least reference) suggestions on how to properly interpret each of the cited common misinterpretations/misuses. For example, give readers an explicit way to interpret a chi-square test result when it rejects the model under study and when it does not, specifically in regard to whether or not the model is, indeed, problematic. Including advice on how this interpretation should change in the presence of large (or small N). What are appropriate steps for the suggested “diagnostic investigations of the various potential reasons for failure”? This type of advice or suggestion is given in some argument sections, but an effort could be made to do this throughout to be more constructive and instructive. I think this can be done while staying on the main point of dissembling the arguments against the
model fit test. Maybe for each argument when/if it comes down to a misinterpretation or mis-step committed by a hypothetical researcher, give some brief guidance on how to avoid this.

I have expanded discussion of appropriate diagnostic investigations that would follow model failure according to χ². I agree that diagnostic direction is helpful but worry that too much extension in this regard would detract from the main message. Diagnostics are a very large topic – too large to be incorporated without muddying the discussion of respecting test evidence.

The author argues against the categorical rejection of the Chi-square test of SEM model fit due to its imperfections.

Not quite. I claim the supposed “imperfections” are usually mis-applications or mis-understanding of the statistics; or are camouflage for the deficient research practice of disregarding potentially important evidence of model misspecification problems.

However, the author then seems to categorically reject all other fit indices as they are imperfect. Did I misinterpret this?

Yes and No. Yes all indices are categorically rejected as an excuse for disregard of, or failure to appropriately report and examine, χ² evidence of problems. No I have not questioned the indices as “imperfect” – in fact I have not even named many of the indices. The indices may have other reasonable uses. It is only the inappropriate use of indices to displace χ² that is relevant here.

If not, this doesn’t seem to be the most reasoned response. The author makes a good point that since no fit indices reliably give an indication of model misspecification, using a CI around one or using several separate indices together doesn’t solve the problem. Each fit index was developed, to some extent, to address problems with other fit indices, so while acknowledging that no fit index is perfect, it would appear that they may provide some additional information when presented together.
This statement refers to the fit indices as potentially useful as INDICES. The article concerns where indices suffice to replace/assist the chi-square TEST, not the chi-square index.

Instead of discouraging the use of CIs or multiple fit indices at once, why not advise that if the Chi-square test is used in conjunction with other model fit indices that the researcher not misinterpret a significant Chi-square test paired with a fit index showing ‘good fit’ as evidence that the Chi-square test is wrong and the fit index is right, but rather an opportunity to provide more information about the model, or as a way to set up the subsequent model fit diagnostics that were performed and resulted in the researcher’s ultimate conclusion about his/her model.

Unfortunately the indices provide minimal diagnostic assistance. But again, discussion of how to attempt to diagnostically determine what, or which multiple things, are wrong/problematic is beyond the current discussion that addresses the issue of what constitutes evidence of problems.

Or selecting a priori fit statistics that will be reported rather than only presenting ones that support the model.

As long as the indices are viewed as supporting or not supporting the model, they are prone to misapplication as being in disagreement with significant $\chi^2$'s. If indices are to be reported, they should be used to describe some feature(s) of the model disconnected from the issue of whether or not the data is consistent with the model’s implications. That would leave $\chi^2$ as the appropriate test of the model, and the indices as describing other model features. Since the indices are commonly used precisely to argue against $\chi^2$, it will take considerable remapping of the index-terrain to institute such a change. “a priori” selection of indices intended to discount or argue against $\chi^2$ would incorporate the same pitting of indices against $\chi^2$ that is fundamentally problematic. What else, or what in addition, each index might contribute is beyond what the current article can consider.

Further, advising that a non-significant Chi-square test paired with ‘good’ fit statistic results does not represent independent evidence of model fit and
should not be presented as such.

Yes, even with $\chi^2$-fitting models the indices should not be seen as further supporting the assessment of whether or not the model is consistent with the current data. The indices will stand-or-fall on the basis of whatever else they are able to report. The issue of what “else” any index might usefully report is quite separate from respecting or disrespecting $\chi^2$ test evidence.

To this same point, it’s strange to me that the author would be for the Chi-square test while at the same time against all other fit statistics as in many cases these results should be related as they are often based on the same log-likelihood function.

The important difference between a fit TEST and fit INDICES remains even when both are based on the same fit function. And some indices assigned “supposedly-acceptable”-cut-values (so they seem like tests) adjust the fit function value inappropriately. For example, the historically-declining “acceptable”-cut-value on $\chi^2$/df is simply statistically inappropriate because the statistically proper way to acknowledge the connection between $\chi^2$ and degrees of freedom is to assess the observed $\chi^2$ against the $\chi^2$ sampling distribution having the appropriate degrees of freedom – not to divide $\chi^2$ by the degrees of freedom. Yes many indices are based on the same fit function. But how the indices are connected to that fit function can discredit the indices. Furthermore, even mathematically acceptable indices can be sabotaged by inappropriate supplemental claims – such as when some non-zero RMSEA can (supposedly but actually only inappropriately) be discounted as “error of approximation”, as indicated in the article.

The reviewer has noted some instances of misinterpretation of references and quotes that should be modified.

I have clarified and emended quotes/references to: Ledermann; Box; and Hu and Bentler (thanks to the substantively-targeted comments of Reviewer 1).

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