Validity and Model Fit

Mike Edwards

The Ohio State University

July 22, 2014

Talk Outline

- Preamble
- Model fit
- Validity
- Validity and model fit

Acknowledgements

I wanted to acknowledge a few individuals who have made contributions to my thinking on these issues:

- Andrew Bodine, EunHee Keum, Yinghao Sun, and Leanne Williamson
- Michael Browne, Li Cai, Bob Cudeck, Bud MacCallum, Albert Maydeu-Olivares, Roger Millsap, Kris Preacher, David Thissen, and R.J. Wirth
- Denny Borsboom, Michael Kane, Keith Markus, and Samuel Messick

Roger Millsap

Millsap (2007) - Structural equation modeling made difficult

Millsap (2013) - A simulation paradigm for evaluating approximate fit

Alternate universes

The title for this talk that appears in the program is: “Some Thoughts on Validity and Model Fit.”

In hindsight, the inclusion of “Some Thoughts” didn’t seem particular helpful, so I dropped it and just called the talk “Validity and Model Fit.”

Since I was changing the title, I toyed with a few other options:

Alternate universes

The title for this talk that appears in the program is: “Some Thoughts on Validity and Model Fit.”

In hindsight, the inclusion of “Some Thoughts” didn’t seem particular helpful, so I dropped it and just called the talk “Validity and Model Fit.”

Since I was changing the title, I toyed with a few other options:

- It’s a good thing there’s no NHST for validity

Alternate universes

The title for this talk that appears in the program is: "Some Thoughts on Validity and Model Fit."

In hindsight, the inclusion of "Some Thoughts" didn't seem particular helpful, so I dropped it and just called the talk "Validity and Model Fit."

Since I was changing the title, I toyed with a few other options:

- It's a good thing there's no NHST for validity
- More rules than reason

Alternate universes

The title for this talk that appears in the program is: “Some Thoughts on Validity and Model Fit.”

In hindsight, the inclusion of “Some Thoughts” didn’t seem particular helpful, so I dropped it and just called the talk “Validity and Model Fit.”

Since I was changing the title, I toyed with a few other options:

- It’s a good thing there’s no NHST for validity
- More rules than reason
- Tenure as a gateway drug to validity research

Alternate universes

The title for this talk that appears in the program is: “Some Thoughts on Validity and Model Fit.”

In hindsight, the inclusion of “Some Thoughts” didn’t seem particular helpful, so I dropped it and just called the talk “Validity and Model Fit.”

Since I was changing the title, I toyed with a few other options:

- It’s a good thing there’s no NHST for validity
- More rules than reason
- Tenure as a gateway drug to validity research
- What are we doing?

Spoiler Alert

TL;DR

Model fit might have a validity problem.

Model fit

Model fit has a long and complex history in structural equation modeling (SEM). In what follows, I'll present a selective overview highlighting changes to our thinking about model fit over time.

I talk about SEM as that's where a lot of the conversation has occured/is occurring, but I think many, if not all, of these points generalize to other models/frameworks.

Model fit - In the beginning...

Early in the history SEM, all we really had was the χ^2 test of exact fit. This test has some functional difficulties in practice (e.g., sensitivity to sample size), but there are more general reasons not to prefer it:

> *In applications of the analysis of covariance structures in the social sciences it is implausible that any model that we use is anything more than an approximation to reality. Since a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless even to try and test whether it is true.*
> *(Browne & Cudeck, 1993, p.137)*

Model fit - The age of creation

Rather than trying to ask whether a model is correct, or fits the population covariance matrix exactly, it is sensible to assess the degree of lack of fit of the model.
(Browne & Cudeck, 1993, p.137)

Model fit - The age of creation

> *Rather than trying to ask whether a model is correct, or fits the population covariance matrix exactly, it is sensible to assess the degree of lack of fit of the model.*
> *(Browne & Cudeck, 1993, p.137)*

And lots of folks got busy developing new ways to assess approximate fit. There are many such tools we have available to us. One of them (at least) had its origin at a talk at this very conference in 1980 (Steiger & Lind, 1980). It's a pretty popular one and has some nice features (e.g., better known statistical properties), so I will occasionally reference it in the remainder of the talk.

However, I don't think anything I'm going to say is unique to the RMSEA, it's just a useful proxy.

Model fit - The age of early use

Lots of researchers wanted to use SEM and before long, reviewers of their submitted manuscripts were asking: “That’s great, but does your model fit?”

Since most models (in my experience) don’t fit using the test of exact fit, there was a real interest among applied users for something else.

Since the approximate fit question isn’t obviously uninteresting a priori, it seems like an improvement.

Model fit - The age of early use

Software began to provide users with lots (and lots) of indexes meant to assess approximate fit. In some cases it is actually quite overwhelming to see the output from modern SEM software in this regard.

Users, after fitting a model to their data, now had a new problem: “My RMSEA is 0.07 - is that good?”

Model fit - The age of early use

Software began to provide users with lots (and lots) of indexes meant to assess approximate fit. In some cases it is actually quite overwhelming to see the output from modern SEM software in this regard.

Users, after fitting a model to their data, now had a new problem: “My RMSEA is 0.07 - is that good?”

Many of the early papers that developed these various measures (or introduced them to a wider audience) provided some form of guidance to readers based on the experience of the authors:

> *Practical experience has made us feel that a value of the RMSEA of about 0.05 or less would indicate a close fit of the model in relation to the degrees of freedom...We are also of the opinion that a value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with a RMSEA greater than 0.1. (Browne & Cudeck, 1993, p.144)*

Model fit - The age of early use

Lest readers get the wrong impressions, early pioneers in approximate fit were very careful to provide very bright, very bold warning signs:

> *This figure is based on subjective judgment. It cannot be regarded as infallible or correct, but it is more reasonable than the requirement of exact fit with the RMSEA=0.0. (Browne & Cudeck, 1993, p.144)*

Model fit - The age of early use

Lest readers get the wrong impressions, early pioneers in approximate fit were very careful to provide very bright, very bold warning signs:

> *This figure is based on subjective judgment. It cannot be regarded as infallible or correct, but it is more reasonable than the requirement of exact fit with the RMSEA=0.0. (Browne & Cudeck, 1993, p.144)*

But our colleagues like dichotomies. And these gently suggested guideposts became iron-clad mandates very quickly.

Model fit - The reformation

Hu and Bentler (1999) approached the problem with a series of simulation studies to see how well these informal (but now extremely pervasive) guidelines actually fared in a few specific conditions.

Model fit - The reformation

Hu and Bentler (1999) approached the problem with a series of simulation studies to see how well these informal (but now extremely pervasive) guidelines actually fared in a few specific conditions.

I have a lot more sympathy for this paper than I used to. You can only tell folks not to do something for so long - if they are going to do it anyway then maybe you can at least give them a safer way to do it. I haven't spoken to the authors about it, but I think I hear a grudging acceptance:

> *Although it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various conditions... (Hu & Bentler, 1999, p.27)*

Model fit - The reformation

Hu and Bentler (1999) approached the problem with a series of simulation studies to see how well these informal (but now extremely pervasive) guidelines actually fared in a few specific conditions.

I have a lot more sympathy for this paper than I used to. You can only tell folks not to do something for so long - if they are going to do it anyway then maybe you can at least give them a safer way to do it. I haven't spoken to the authors about it, but I think I hear a grudging acceptance:

> *Although it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various conditions... (Hu & Bentler, 1999, p.27)*

A quick check on Web of Science over the weekend showed that it has been cited over 10,000 times.

Model fit - The counter-reformation

Then a lot of conversation ensued about Hu and Bentler (1999). These took on various flavors:

Model fit - The counter-reformation

Then a lot of conversation ensued about Hu and Bentler (1999). These took on various flavors:

- None of this approximate fit stuff works: Hayduk and Glaser (2000), Barrett (2007)

Model fit - The counter-reformation

Then a lot of conversation ensued about Hu and Bentler (1999). These took on various flavors:

- None of this approximate fit stuff works: Hayduk and Glaser (2000), Barrett (2007)
- Issues with specific methodology: Marsh et al. (2004), Fan and Sivo (2005), Preacher and Merkle (2012), Savalei (2012)

Model fit - The counter-reformation

Then a lot of conversation ensued about Hu and Bentler (1999). These took on various flavors:

- None of this approximate fit stuff works: Hayduk and Glaser (2000), Barrett (2007)
- Issues with specific methodology: Marsh et al. (2004), Fan and Sivo (2005), Preacher and Merkle (2012), Savalei (2012)
- Issues with the question: Nye and Drasgow (2011), Preacher et al. (2013)

Model fit - The counter-reformation

Then a lot of conversation ensued about Hu and Bentler (1999). These took on various flavors:

- None of this approximate fit stuff works: Hayduk and Glaser (2000), Barrett (2007)
- Issues with specific methodology: Marsh et al. (2004), Fan and Sivo (2005), Preacher and Merkle (2012), Savalei (2012)
- Issues with the question: Nye and Drasgow (2011), Preacher et al. (2013)
- Create your own cut-point: Millsap (2013)

Now for something completely different

Obviously the discussion of fit is vastly more complicated than this, but hopefully I've captured some of the macro issues and conveyed a general sense of my reading of the development in the literature.

Next we'll talk a bit about modern understandings of validity both at a theoretical level and in practice.

If you want to be "caught up" on validity, read Messick (1989), Kane (2013), and Markus and Borsboom (2013)

Now for something completely different

Obviously the discussion of fit is vastly more complicated than this, but hopefully I've captured some of the macro issues and conveyed a general sense of my reading of the development in the literature.

Next we'll talk a bit about modern understandings of validity both at a theoretical level and in practice.

If you want to be "caught up" on validity, read Messick (1989), Kane (2013), and Markus and Borsboom (2013)

> *In contrast to test theory as a whole, test validity represents the least mathematical specialization within the most mathematical sub-field of less mathematical disciplines.*
> *(Markus & Borsboom, 2013, p.xiii)*

Validity

The dominant definition of validity, at least in the literature, is:

> *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.*
> *(emphasis in original, Messick, 1989, p.13)*

Validity

The dominant definition of validity, at least in the literature, is:

> *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.*
> *(emphasis in original, Messick, 1989, p.13)*

But a competing definition, which perhaps more closely matches individuals' mental models of validity is:

> *A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes.*
> *(Borsboom et al., 2004, p.1061)*

The practice of validation

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on those scores...Validation can then be though of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions.
(Kane, 2013, p.1)

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses.
(AERA, APA, & NCME, 1999, p.13)

A thought begins to form...

More broadly, we are concerned with the validity of everything we use, and not just the validity of all the measurement procedures used, but also the validity of the research design, the validity of the experimental methods (including the validity of the stimuli themselves), and the validity of our conclusions and inferences. (Fiske, 2002, p.176)

What do fit measures and test scores have in common?

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on model fit.

Which leads me to a question: What kind of inferences do we make (or want to make) when a model fits well?

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.
- ...this is a useful model.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.
- ...this is a useful model.
- ...this model will replicate.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.
- ...this is a useful model.
- ...this model will replicate.
- ...our theory is correct.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.
- ...this is a useful model.
- ...this model will replicate.
- ...our theory is correct.
- ...our theory is plausible.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ...this is the generating model.
- ...this is a useful model.
- ...this model will replicate.
- ...our theory is correct.
- ...our theory is plausible.
- ...we can proceed with further evaluation of our model.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ~~...this is the generating model.~~
- ...this is a useful model.
- ...this model will replicate.
- ...our theory is correct.
- ...our theory is plausible.
- ...we can proceed with further evaluation of our model.

My model fits well, so...

I did a little thought experiment by asking the question to myself and by reading published applications and trying to see what I think the authors' answer would be. Here's what I came up with.

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.
- ~~...this is the generating model.~~
- ...this is a useful model.
- ...this model will replicate.
- ~~...our theory is correct.~~
- ...our theory is plausible.
- ...we can proceed with further evaluation of our model.

Breaking down possible inferences

My model fits well, so...

- ...we can stop adding correlated residuals.
- ...we can (probably) get this published.

While these may be true, they aren't exactly at the level one would hope for scientific discourse ala Popper or Kuhn, are they?

Breaking down possible inferences

My model fits well, so...

- ...this is a useful model.

Breaking down possible inferences

My model fits well, so...

- ...this is a useful model.

  Given that a proposed model does not provide an exact fit, an approximate fit index will summarize the degree of misfit. The tacit rationale for such indices is that the degree of misfit is relevant information when deciding whether the model is scientifically useful. (Millsap, 2007, p.876)

Breaking down possible inferences

My model fits well, so...

- ...this is a useful model.

  Given that a proposed model does not provide an exact fit, an approximate fit index will summarize the degree of misfit. The tacit rationale for such indices is that the degree of misfit is relevant information when deciding whether the model is scientifically useful. (Millsap, 2007, p.876)

  Fit indices should not be regarded as measures of usefulness of a model. They contain some information about the lack of fit of a model, but none about plausibility.
  (Browne & Cudeck, 1993, p.157)

Breaking down possible inferences

My model fits well, so...

- ...this model will replicate.

There are specific fit measures meant to assess this kind of thing and there is some work, for example Preacher et al. (2013), that suggests some of the usual suspects in the model fit space are useful for this goal.

Breaking down possible inferences

My model fits well, so...

- ...our theory is plausible.

Breaking down possible inferences

My model fits well, so...

- ...our theory is plausible.

  However, if a theory does not constrain possible outcomes, the fit is meaningless.
  (Roberts & Pashler, 2000, p.359)

Breaking down possible inferences

My model fits well, so...

- ...our theory is plausible.

  However, if a theory does not constrain possible outcomes, the fit is meaningless.
  (Roberts & Pashler, 2000, p.359)

  At most, one can conclude that a well-fitting model is one plausible representation of the underlying structure from a larger pool of plausible models.
  (Tomarken & Waller, 2003, p.580)

Breaking down possible inferences

My model fits well, so...

- ...we can proceed with further evaluation of our model.

Breaking down possible inferences

My model fits well, so...

- ...we can proceed with further evaluation of our model.

  The goal of model selection in structural equation modeling (SEM) is to find a useful approximating model that (a) fits well, (b) has easily interpretable parameters, (c) approximates reality in as parsimonious a fashion as possible, and (d) can be used as a basis for inference and prediction.
  (Preacher & Merkle, 2012, p.1)

Breaking down possible inferences

My model fits well, so...

- ...we can proceed with further evaluation of our model.

  The goal of model selection in structural equation modeling (SEM) is to find a useful approximating model that (a) fits well, (b) has easily interpretable parameters, (c) approximates reality in as parsimonious a fashion as possible, and (d) can be used as a basis for inference and prediction.
  (Preacher & Merkle, 2012, p.1)

  These examples prove that there is no direct or functional relationship between degree of model misspecification and degree of approximate fit, but a functional relationship is not strictly needed. What is needed is some level of relationship that would support continued use of such indices. Most investigators assume that such a relationship exists, but the question has received surprisingly little direct study.
  (Millsap, 2007, p.876)

Next steps

It seems to me, at this point, that we have some work to do to shore up the valid use of model fit measures. From what I've seen, we have not been very clear to users about what inferences are possible and perhaps have not done as good a job as we could have with describing where model fit resides in the process of model comparison/selection.

Next steps

It seems to me, at this point, that we have some work to do to shore up the valid use of model fit measures. From what I've seen, we have not been very clear to users about what inferences are possible and perhaps have not done as good a job as we could have with describing where model fit resides in the process of model comparison/selection.

I think there are good cases to be made for some of the inferences I've discussed here, but the tenuous link between model fit and model utility has been genuinely perplexing to me.

Next steps

It seems to me, at this point, that we have some work to do to shore up the valid use of model fit measures. From what I've seen, we have not been very clear to users about what inferences are possible and perhaps have not done as good a job as we could have with describing where model fit resides in the process of model comparison/selection.

I think there are good cases to be made for some of the inferences I've discussed here, but the tenuous link between model fit and model utility has been genuinely perplexing to me.

I also think that, what we would like to be able to say is some version of: A model that fits well has a higher probability of being scientifically useful than one that does not.

Next steps

It seems to me, at this point, that we have some work to do to shore up the valid use of model fit measures. From what I've seen, we have not been very clear to users about what inferences are possible and perhaps have not done as good a job as we could have with describing where model fit resides in the process of model comparison/selection.

I think there are good cases to be made for some of the inferences I've discussed here, but the tenuous link between model fit and model utility has been genuinely perplexing to me.

I also think that, what we would like to be able to say is some version of: A model that fits well has a higher probability of being scientifically useful than one that does not. *(But of course even this is context dependent...)*

Parting thoughts

Parting thoughts

The interpretability of a model can be judged only subjectively and is not amenable to the application of statistical methods. This does not render this characteristic of a model any less important; it is only more difficult to investigate.
(Browne & Cudeck, 1993, p.136)

Parting thoughts

The interpretability of a model can be judged only subjectively and is not amenable to the application of statistical methods. This does not render this characteristic of a model any less important; it is only more difficult to investigate.
(Browne & Cudeck, 1993, p.136)

Who said structural equation modeling was easy?
(Millsap, 2007, p.880)

The end

Thanks

edwards.134@osu.edu

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061-1071.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (p. 136-162). Newbury Park, CA: Sage.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*, 343-367.

Fiske, D. W. (2002). Validity for what? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (p. 169-178). Mahwah, NJ: Lawrence Erlbaum Associates.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling*, *7*, 1-35.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement: Third edition* (p. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*, 875-881.

Millsap, R. E. (2013). A simulation paradigm for evaluating approximate fit. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (p. 165-182). New York, NY: Routledge.

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*, 548-570.

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, *17*, 1-14.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspetive. *Multivariate Behavioral Research*, *48*, 28-56.

Roberts, S., & Pashler, H. (2000). How persuasive is good fit? A comment on theory testing. *Psychological Review*, *107*, 358-367.

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, *72*, 910-932.

Steiger, J. H., & Lind, J. (1980). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society. Iowa City, IA.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well fitting” models. *Journal of Abnormal Psychology*, *112*, 578-598.