



If Your Alpha Coefficient is “Flashing Red,” Check Your Model!

Randall Bender, PhD, Senior Psychometric Statistician, Outcomes Research, Evidera

The old marketing slogan, “One size fits all,” has never been entirely true. Many of us confirm this every time we go shopping for clothes. The same reality confronts us in patient-reported outcome (PRO) scale development and scale assessment. One model does not fit all data or scale types, which can have serious ramifications for researchers dependent on good scales. By applying the wrong model, scales can be improperly assessed and erroneously dismissed, wasting research time and dollars.

In scale development, the reflective indicator model (RIM) underlies most scaling methodologies, from coefficient alpha to factor analysis and item response theory (IRT)/Rasch modeling. The RIM treats the observed variables as reflections of underlying latent variables or true scores. Unfortunately, this model is indiscriminately applied to many datasets given its dominance. Given the methods that have been developed, early scaling methodologists clearly must have dealt primarily with the type of data best served by this model. One sees little reference to alternative models in much of the classical literature, leading Bollen¹ to make the startling note that the first systematic discussion of model selection occurred relatively late in a 1971 paper by Blalock.²

Today, in the face of an explosion of scale creation, methodologists are facing a greater diversity of scales as scientific measurement moves into new areas and applications. Today,

methodologists are recognizing that the RIM, the cornerstone of classical and even most modern scale development, is not appropriate for some types of scale data and that alternative models and assessments need to be developed and appropriated. Recognizing the need for alternative models is important as perfectly good scales may be discarded if they do not meet expected measurement standards (e.g., coefficient alpha criteria, model fit and proper parameter estimates for factor analysis and IRT). This article will highlight one important alternative model to raise awareness of the importance of choosing the correct model for scale assessment.

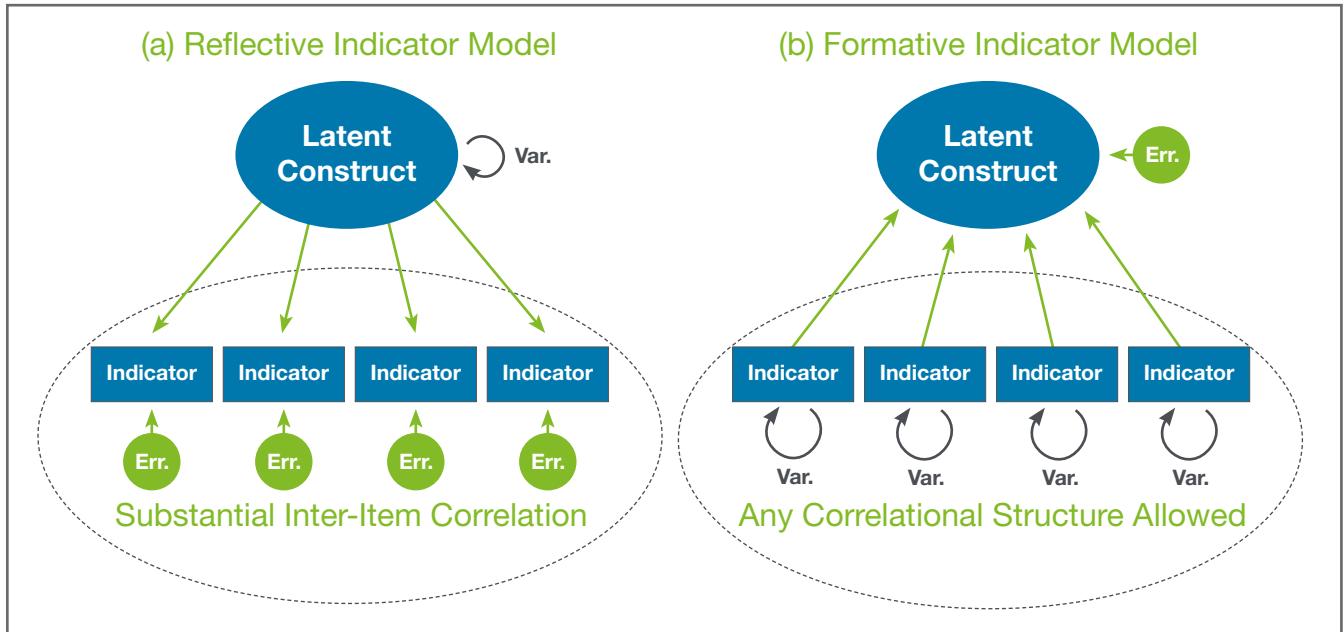
To appreciate the difference between scale models, one must understand the hypothesized relationships in each model between the observed variables and the latent variable of true interest. The RIM is defined by its assumption that each indicator *reflects* the state of the latent variable, such that if that latent variable changes, every connected indicator should probabilistically “reflect” this by realizing some particular change. Of course the “reflection” may be imperfect as if by a carnival mirror because of measurement error. Another name used in the literature for such an indicator is an “effect” indicator because it shows the *effect* of the latent construct. Most of our psychometric methodologies assume there is a common source of variance for the observed variables and that this common source of variance is provided

by the latent variable varying over individuals, causing correlation across individuals in the observed variables.

A less common but important alternative model to be considered, which is more appropriate for some scales, is called a formative indicator model (FIM). In this model, the causal relationship between the observed variables and the latent construct is reversed. The measured or observed variables in this case *construct or form* the latent variable (hence the term “formative”), which is in effect assembled from the items. Another name for this type of model is a causal indicator model because the indicator *causes* the latent construct. *Figure 1* displays a visual representation of the two models, (a) a RIM and (b) a FIM, displaying the key difference between them lying in the direction of influence of the arrows connecting the latent construct in the ovals with the four indicators, indicated by square boxes. The figure also signals a less obvious potential difference between the two, namely the degree to which there is inter-item correlation. More discussion of these two models can be found in Bollen and Lennox.³ The FIM is also discussed in detail in Bollen and Bauldry⁴ along with a third model not presented here.

Before discussing the problems with applying typical psychometric methods to FIM scales, a few examples are in order. A commonly encountered example of an FIM scale type is the typical stress scale, in which a list of stressors

Figure 1: Two different models underlying the most commonly used scales



is presented and the respondent indicates whether that particular stress-inducing event has occurred in that person's life during the stated period of time. The guiding theory posits that occurrence of such events would likely raise that person's stress level. What is important to note with this stress scale example is the relationship between the variable of interest, the person's stress level, and the observed variables, the individual stressors. The occurrence of the observed stressor has a causal effect on the unobserved stress level. The reverse, a manipulation of the stress level, would not cause the occurrence of each of the stressful events.

Another example might be a social engagement scale. Here individual items, time spent with family, time spent with friends, time spent with work colleagues, etc., together constitute an overall social engagement, but each bundle of engagement time builds separately on the others to form the overall engagement variable. It does not make sense to vary the overall social engagement without deterministically (not probabilistically) varying at least one of the individual components. However, there may be another

latent variable, say a latent sociability variable that could drive each of those parts. This example highlights the fact that carefully thinking about the latent variable and any causal direction vis-à-vis observed variables is crucial, as the same set of observed variables can represent two different latent variables depending on how they are modeled. While an individual sociability characteristic or trait may be related to a social engagement construct, they are clearly not the same variable.

A third example, seen in the outcomes research field may be in the assessment of symptoms. Such assessments may be used in a symptom impact index, designed to measure the cumulative impact of the person's symptom experience on his or her health-related quality of life. In this example, the best model is an FIM, as the symptom experiences add up to and are causal of an overall symptom impact. An alternative use of symptom indices occurs in measures of disease severity, wherein symptom expression is an indicator of how severe the person's disease state is. For this use, the RIM is appropriate as the observed symptoms are seen as reflective of the underlying disease

severity. Again, as in the previous example, depending on the causal direction assumed in the measurement model used, the same set of observed variables may be used for two different latent variables. Sometimes more refined measurement is obtained by using item wording that focuses respondent attention to symptom *impact* on health-related quality of life, so the question is not just about the presence of the symptom, but about the degree to which its presence is having an impact on daily life.

A corollary of the causal direction embedded in each model is the correlational structure and item independence. In the RIM, a change in one observed variable should be accompanied by changes in all the variables as the implication of the model is that the latent variable must have changed with the observed variable since it is but a reflection of the latent variable. In the FIM, any observed variable can change independently, not necessitating a correlated change in any other observed variable per the model. The degree of correlation among the observed variables in the FIM can vary from high to none at all.

The complete lack of any specification of inter-item correlation among the observed indicators in the FIM is the reason why FIM scales may, on occasion, meet good scale criteria, but more often will fail to meet such criteria; this is where researchers can encounter difficulties in their scale development if they use the wrong model. In RIM scales, there is a strong basis for correlation among the indicators because all of them share a common cause.


In contrast, the FIM scale contains no common cause of the indicators, so there is nothing in the model that specifies any necessary degree of correlation. Completely uncorrelated items may still form a very good formative index. For example, in a stress index, the two items, (a) being the victim of an automobile accident and (b) having a close family member who is terminally ill, may have virtually no correlation. There is no model assumption that raises the likelihood of both circumstances happening at the same time. Components of scale analysis that assume a common correlation among all the indicators, and test for or assess it, are quite appropriate for the RIM, but not appropriate for the FIM. Coefficient alpha assesses common inter-item correlation. Factor analysis estimates parameters around an assumed

common cause of observed variable correlation. (*Figure 1a* is the classic graphical presentation of the basic factor analysis model.) Similarly with a slightly different model, IRT and Rasch models are built around a common source (the latent trait) of item correlation (response propensity).

When these scale analysis methods are applied to formative scales they occasionally will, but more likely will not, meet certain required criteria. It all depends on how much correlation exists among the formative indicators, either from other common causes some of the items may share or due to causal relationships among the indicators themselves. When formative scale items do show considerable correlation and the typical psychometric analysis is used with this data, this correlation may mask FIM items as RIM items, wrongly attributing that observed inter-item correlation to a latent construct, which is assumed under RIM to be a “causal” agent. The unfortunate consequence is that when formative items are tested with reflective model tests, they either (1) provide deceptive information in the form of parameter estimates for a completely miss-specified model, or (2) when inter-item correlation is low or non-existent (which is entirely acceptable in the formative

WHEN INAPPROPRIATE MODEL ASSUMPTIONS ARE APPLIED, PROBLEMS WILL BE ENCOUNTERED AND RESEARCHERS WILL BE PLAGUED BY PUZZLING AND INCONSISTENT RESULTS.

scale), they may fail to meet the required levels of correlation and be inappropriately discarded. For further information regarding formative scales and their assessment, see Bollen and Ting,^{5,6} Hipp et al.,⁷ and MacCallum and Browne.⁸

It is a very real possibility that researchers today may encounter and need to assess such scales. (Some may even have a mix of reflective and formative indicators.) By starting with an awareness of this issue and thus being able to make appropriate model choices, scale analysis can proceed in a sensible way. When inappropriate model assumptions are applied, problems will be encountered and researchers will be plagued by puzzling and inconsistent results. 

For more information, please contact Randall.Bender@evidera.com.

References

- ¹ Bollen KA. Interpretational Confounding is Due to Misspecification, Not to Type of Indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychol Methods*. 2007 Jun; 12(2): 219-228.
- ² Blalock HM, Jr. Causal Models Involving Unmeasured Variables in Stimulus-response Situations. 1971; pp. 335-347 in H. M. Blalock, Jr. (ed.) *Causal Models in the Social Sciences*. New York: Aldine-Atherton.
- ³ Bollen KA, Lennox R. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychol. Bull.* 1991 Sep; 110(2):305-314.
- ⁴ Bollen KA, Bauldry S. Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. *Psychol Methods*. 2011 Sep; 16(32):265-284.
- ⁵ Bollen K, Ting K. Confirmatory Tetrad Analysis. In: Marsden PM, ed. *Sociological Methodology*. Cambridge, MA: Wiley-Blackwell; 1993:147-175.
- ⁶ Bollen KA, Ting KF. A Tetrad Test for Causal Indicators. *Psychol Methods*. 2000 Mar; 5(1):3-22.
- ⁷ Hipp JR, Bauer DJ, and Bollen KA. Conducting Tetrad Tests of Model Fit and Contrasts of Tetrad-Nested Models: A New SAS Macro. *Struct Equ Modeling*. 2005; 12(1):76-93.
- ⁸ MacCallum RC, Browne MW. The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues. *Psychol Bull.* 1993 Nov; 114(3):533-541.