**Literature advanced topics**

**Bollen & Pearl (2013). Eight myths about causality and structural equation models**

Causality was at the center of the early history of *structural equation models (SEMs),* which continues to serve as the most popular approach to causal analysis in the social sciences. Through decades of development, critics and defenses of the capability of SEMs to support causal inference have accumulated. A variety of misunderstandings and myths about the nature of SEMs and their role in causal analysis have emerged, and their repetition has led some to believe they are true. Our chapter is organized by presenting eight myths about causality and SEMs in the hope that this will lead to a more accurate understanding. More specifically, the eight myths are the following: (1) SEMs aim to establish causal relations from associations alone, (2) SEMs and regression are essentially equivalent, (3) no causation without manipulation, (4) SEMs are not equipped to handle nonlinear causal relationships, (5) a potential outcome framework is more principled than SEMs, (6) SEMs are not applicable to experiments with randomized treatments, (7) mediation analysis in SEMs is inherently non-causal, and (8) SEMs do not test any major part of the theory against the data. We present the facts that dispel these myths, describe what SEMs can and cannot do, and briefly present our critique of current practice using SEMs. We conclude that the current capabilities of SEMs to formalize and implement causal inference tasks are indispensible; its potential is even greater.

**Model and assumptions of SEMs**

This SEM consists of two major parts. The first is a set of equations that give the causal relations between the substantive variables of interest, also called “latent variables,” because they are often inaccessible to direct measurement (e.g., self-esteem). The second part of the model ties the observed variables or measures to the substantive latent variables in a two-equation measurement model. The SEM explicitly recognizes that the substantive variables are likely measured with error and possibly measured by multiple indicators. Therefore, the preceding separate specification links the observed variables that serve as indicators to their corresponding latent variables.

**Eight myths**

*1. SEMs aim to establish causal relations from associations alone*

SEM does not aim to establish causal relations from associations alone. Perhaps the best way to make this point clear is to state formally and unambiguously what SEM does aim to establish. SEM is an inference engine that takes in two inputs, qualitative causal assumptions and empirical data, and derives two logical consequences of these inputs: quantitative causal conclusions and statistical measures of fit for the testable implications of the assumptions. Failure to fit the data casts doubt on the strong causal assumptions of zero coefficients or zero covariances and guides the researcher to diagnose or repair the structural misspecifications. Fitting the data does not “prove” the causal assumptions, but it makes them tentatively more plausible. Any such positive results need to be replicated and to withstand the criticisms of researchers who suggest other models for the same data.

*2. SEM and regression are essentially the same*

“In a structural equation model each equation represents a causal link rather than a mere empirical association. In a regression model, on the other hand, each equation represents the conditional mean of a dependent variable as a function of explanatory variables.” Admittedly, Goldberger’s quote emphasizes the weak causal assumptions over the strong causal assumptions as distinguished by us earlier, but it does point to the semantic difference between the coefficients originating with a regression where no causal assumptions are made versus from a structural equation that makes strong and weak causal assumptions. The preoccupation of early SEM researchers with the identification problem testifies to the fact that they were well aware of the causal assumptions that enter their models and the acute sensitivity of SEM claims to the plausibility of those assumptions.

*3. No causation without manipulation*

A softer view of the “no causation without manipulation” motto is that actual physical manipulation is not required. Rather, it requires that we be able to imagine such manipulation. A SEM specification incorporates the causal assumptions of the researcher.

*4. The potential outcome framework is more principled than SEMs*

Notwithstanding these critics, a productive symbiosis has emerged that combines the best features of the two approaches (Pearl 2010). It is based on encoding causal assumptions in the transparent language of (nonparametric) SEM, translating these assumptions into counterfactual notation, and then giving the analyst an option of either pursuing the analysis algebraically in the calculus of counterfactuals or use the inferential machinery of graphical models to derive conclusions concerning identification, estimation, and testable implications.

*5. SEMs are not equipped to handle non-linear causal relationships*

The SEM presented so far is indeed linear in variables and in the parameters. We can generalize the model in several ways. First, there is a fair amount of work on including interactions and quadratics of the latent variables into the model. These models stay linear in the parameters, though they are nonlinear in the variables. Another nonlinear model arises when the endogenous observed variables are not continuous.

*6. SEMs are less applicable to RCTs*

This misunderstanding is not as widespread as the previous ones. However, the heavy application of SEMs to observational (nonexperimental) data and its relative infrequent use in randomized experiments have led to the impression that there is little to gain from using SEMs with experimental data. Drawing on these sources, we summarize valuable aspects of applying SEMs to experiments. In brief, SEMs provide a useful tool to help to determine (1) if the randomized stimulus actually affects the intended variable (“manipulation check”), (2) if the output measure is good enough to detect an effect, (3) if the hypothesized mediating variables serve as the mechanism between the stimulus and effect, and (4) if other mechanisms, possibly confounding ones, link the stimulus and effect. These tasks require assumptions, and SEM’s power lies in making these assumptions formal and transparent.

*7. SEM is not appropriate for mediation analysis*

Mediation analysis aims to uncover causal pathways along which changes are transmitted from causes to effects. The myth that SEM is not appropriate for mediation analysis is somewhat ironic in that much of the development of mediation analysis occurred in the SEM literature. In short, SEM largely originated mediation analysis, and it remains at its core.

*8. SEMs do not test any major part of the theory against the data*

In a frequently cited critique of path analysis, Freedman (1987: 112) argues that “path analysis does not derive the causal theory from the data, or test any major part of it against the data.”12 This statement is both vacuous and complimentary. It is vacuous in that no analysis in the world can derive the causal theory from nonexperimental data; it is complimentary because SEMs test *all* the testable implications of the theory, and no analysis can do better. Models typically differ in their empirical implications, but if the empirical implications do not hold, then we reject the model. The causal assumptions are the basis for the construction of the model. Therefore, a rejection of the model means a rejection of at least one causal assumption. It is not always clear which causal assumptions lead to rejection, but we do know that at least one is false and can find the minimal set of suspect culprits. The causal assumptions perpetually remain only a study away from rejection, but the longer they survive a variety of tests in different samples and under different contexts, the more plausible they become. The SEM literature has developed a variety of global (i.e., likelihood ratio test an chi-square test) and local tests that can lead to the rejection of causal assumptions.

**Marsch et al. (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.**

Goodness-of-fit (GOF) indexes provide “*rules of thumb*”—recommended cutoff values for assessing fit in SEM. Hu and Bentler (1999) proposed a more rigorous approach to evaluating decision rules based on GOF indexes and, on this basis, proposed new and more stringent cutoff values for many indexes. This article discusses potential problems underlying the hypothesis-testing rationale of their research, which is more appropriate to testing statistical significance than evaluating GOF. Many of their misspecified models resulted in a fit that should have been deemed acceptable according to even their new, more demanding criteria. Hence, rejection of these acceptable-misspecified models should have constituted a Type 1 error (incorrect rejection of an “acceptable” model), leading to the seemingly paradoxical results whereby the probability of correctly rejecting misspecified models decreased substantially with increasing *N.* In contrast to the application of cutoff values to evaluate each solution in isolation, all the GOF indexes were more effective at identifying differences in misspecification based on nested models. Whereas Hu and Bentler (1999) offered cautions about the use of GOF indexes, current practice seems to have incorporated their new guidelines without sufficient attention to the limitations noted by Hu and Bentler (1999).

**Introduction**

In psychology, data interpretations and their defense is a subjective undertaking that requires researchers to immerse themselves in their data. An important reason for the popularity of goodness-of-fit (GOF) indexes to assess the fit of models in covariance structure analyses is their elusive promise of golden rules—absolute cutoff values that allow researchers to decide whether or not a model adequately fits the data—that have broad generality across different conditions and sample sizes. The impact of Hu and Bentler’s research was twofold. First, they provided an apparently much stronger empirical basis for evaluating the validity of decisions based on cutoff values. Second, their research is apparently leading to the routine use of much more stringent cutoff values for what constituted an acceptable fit. Our comments are directed primarily toward potential problems apparently being incorporated into current practice without appropriate caveats rather than the highly influential and heuristic research by Hu and Bentler (1999) and their more cautious recommendations.

**Rationales for establishing GOF cutoff values**

How good is good enough? There is an implicit assumption in GOF research that sufficiently high levels of GOF (higher than a prescribed cutoff value) are necessary to establish the validity of interpretations of a model. Clearly, a high GOF is not a sufficient basis to establish the validity of interpretations based on the theory underlying the posited model. For example, if theory predicts that a path coefficient should be positive, whereas the observed results show that it is negative, high levels of GOF are not sufficient to argue for the validity of predictions based on the model.

A critical question is whether there are absolute criteria—golden rules or even recommended guidelines—of acceptable levels of fit that are a necessary basis for valid interpretations. Applied researchers seemed to have accepted there are no absolute guidelines for what constitutes necessary levels of reliability accepting imprecise rules of thumb that have limited generality rather than applying golden rules that provide absolute guidance for all situations.

*A normed-reference approach to acceptable levels of GOF*

Analogous to the situation in reliability, rules of thumb about acceptable levels of GOF (e.g., incremental fit indexes > .9) have traditionally been ambit claims based on intuition and accepted wisdom. In summary, based on a normative reference approach to establishing cutoff values on the basis of GOF indexes achieved in current practice, there is some evidence to suggest that even the old cutoff values (e.g., RNI and TLI > .90) are overly demanding in relation to a normative criterion of appropriateness based on the best existing psychological instruments.

**>>Hu and Bentler’s studies on new cutoff criteria for fit indexes**

Al- though one might want a decision rule to be reasonably consistent across *N* for acceptance of the true model, the true model was incorrectly rejected 25%, 3.6%, and 0.1% of the time at small, medium, and large *N,* respectively. Perhaps higher Type 1 error rates would be acceptable at small *N* to minimize Type 2 error rates, although this is inconsistent with traditional approaches to hypothesis testing. Clearly, however, it is desirable for misspecified, false models to be rejected more frequently when *N* increases; but the actual pattern of results reported by Hu and Bentler (1998, 1999) was exactly the opposite. Hence, according to the recommended decision rule for RNI, the likelihood of correctly rejecting a false model was substantial for small *N* and very small for large *N.*

As clearly in an example, the resolution of this apparent paradox is that the population GOF value for the so-called misspecified model actually falls in the acceptance region—RNI > .95—so that, perhaps, this acceptable-misspecified model should be considered an acceptable model rather than a misspecified, false model. This illustration also demonstrates that the conclusions about recommended cutoff values are highly dependent on the particular misspecified models that are used. In contrast, if Hu and Bentler (1998, 1999) chose models of sufficiently extreme levels of misspecification, according to their rationales and evaluation methods, even modest cutoff values (e.g., RNI =.80) would have been highly accurate in discriminating between true and misspecified models and would have led to less stringent criteria of acceptable fits. In addition, a complicated interaction between acceptability or unacceptability of the misspecified models, the data structure (simple vs. complex), the particular index, and sample size demonstrates that broad generalizations based on these data may be unwarranted.

*Summary*. In summary, whereas the behavior of decision rules based on the SRMR is logical in relation to a detailed evaluation of the simulation results, the pattern of results is complicated and apparently different from the RNI. The most dramatic difference between the two is that RNIs were more acceptable for misspecified models based on the simple structure, whereas the SRMRs were more acceptable for misspecified models based on the complex structure. Interpretations were also complicated in that RNI was relatively unbiased by *N,* whereas there was a clear sample size bias in SRMR values. Nevertheless, there was an important consistency in the two sets of results that is the central feature of our comment illustrated in Figure 1. For unacceptable-misspecified models, the behavior of decision rules was reasonable in that rejection rates increased systematically with increasing *N* and were 100% for sufficiently large *N.* However, for acceptable-misspecified models, the behavior of decision rules was apparently inappropriate for a misspecified model in that rejections decreased systematically with increasing *N* and were 0% for sufficiently large *N.*

**Summary**

There are important logical problems underlying the rationale of a hypothesis-testing approach to setting cutoff values for fit indexes. The intent of the GOF indexes has been to provide an alternative to traditional hypothesis-testing approaches based on traditional test statistics (e.g., ML chi-square). However, Hu and Bentler (1998, 1999) specifically evaluated the GOF indexes in relation to a traditional hypothesis-testing paradigm in which the ML chi-square test outperformed all of the GOF indexes. More important, many of the misspecified models considered by Hu and Bentler (1998, 1999) provided a sufficiently good fit in relation to population approximation discrepancy that they should have been classified as acceptable models even according to the more stringent cutoff values proposed by Hu and Bentler (1998, 1999). Hence, rejection of these acceptable models, perhaps, should have constituted a Type 1 error (incorrectly rejecting an acceptable model), further complicating the interpretations of their results. This led to apparently paradoxical behavior patterns in their decision rules. In particular, for most indexes, the probability of correctly rejecting misspecified models systematically decreased with increasing *N,* whereas an appropriate decision rule should result in higher rejection rates for misspecified models with increasing *N.* Particularly because of this heavy reliance on acceptable-misspecified models, the results by Hu and Bentler (1998, 1999) may have limited generalizability to the levels of misspecification experienced in typical practice. Hence, we strongly encourage researchers, textbook authors, reviewers, and journal editors *not to overgeneralize the Hu and Bentler (1998, 1999) results*, transforming heuristic findings based on a very limited sample of misspecified models into golden rules of fit that are broadly applied without the cautions recommended by Hu and Bentler (1999). In contrast to decisions based on comparisons with a priori cutoff values, all of the indexes seemed to be more effective at identifying differences in misspecification based on a comparison of nested models.