DOI 10.25336/csp29397
Canadian Studies in Population 45
no. 3–4 (2018), p. 154–78

Review essay on Rex B. Kline's *Principles and Practice of Structural Equation Modeling*:[1] Encouraging a fifth edition

Leslie Hayduk[2]

Introduction

Kline's fourth edition is reasonably strong but improvable. The text aims to introduce newcomers to fundamental structural equation modeling (SEM) principles, but tends to confuse "Principles" with "Rules." Rules having insufficient grounding in principles leave readers ill-prepared for understanding and responding to changes in previously traditional "rules"—such as those concerning model testing, and latents having single indicators. SEM's foundations would be clearer if Kline began by presenting structural equation models as striving to represent causal effects—a commitment that differentiates structural equation models from regression and encourages model testing. I begin this review by summarizing the covariance/correlation implications of three simple causal structures, which pinpoints multiple text improvements and underpins the discussions of measurement and model testing that follow. Causal structuring also grounds my later comments regarding modelling means/intercepts and interactions. A file of Supplement Sections expands on several points and lists multiple editorial corrections you might pencil into your copy of Kline's text.

Kline's fourth edition is more than one hundred pages longer than his third edition, and is effectively and compactly written. The material has been substantially reorganized, with the most substantive extension being a new chapter on "Graph theory and the structural causal model." The publisher's website contains syntax and output produced by an impressive variety of programs. Kline's detailed discussion of several examples is noteworthy, and I count it as a strong positive that Kline's examples include problems: "not all applications of SEM described in this book are picture perfect, but neither are actual research problems" (p. 1). We will encounter additional "problems" but if Kline prepares a fifth edition, I would recommend retaining the extra-problematic examples, along with supplemental discussions of what led to these slips, and instruction on ways to avoid similar slips. Overall, Kline's text is solid enough to be worth improving. Regrettably, the tone of this review is more negative than I would prefer, but I could not find a way to detail the book's positive features without also indicating some serious concerns. I asked Frank Trovato, editor of *Canadian Studies in Population*, to offer Rex Kline an opportunity to respond to my comments in hope that we might hear of Kline's intentions regarding a fifth edition. I expect other readers of Kline's fourth edition would appreciate your placing a reference to this review (and Kline's response) in whatever copies you encounter. Indeed, my comments presume that you have access to the fourth edition for reference/comparison.

[1] Rex B. Kline. *Principles and Practice of Structural Equation Modeling*. New York: The Guilford Press, 2016. ISBN 978-1-4625-2334-4. Softcover US$65, 534 pp.

[2] Department of Sociology, 5-25 HM Tory Bldg., University of Alberta, Edmonton, AB T6G 2H4; email: lhayduk@ualberta.ca.

The unavoidable implications of three simple causal models

If just variable X linearly causes Y with effect b, as in Figure 1A, this corresponds to the equation

$$Y = a + bX \tag{1}$$

and demands

$$\bar{Y} = a + b\bar{X} \tag{2}$$

$$Var(Y) = b^2 Var(X) \tag{3}$$

$$Cov(XY) = b\,Var(X). \tag{4}$$

The causal world makes variance in the causal variable X (namely, $Var(X)$) produce, and thereby explain, variance in the effect ($Var(Y)$). And the causal world makes variations in one variable (X) produce coordination, correlation, or covariance ($Cov(XY)$) between the causal variable and the effect. Variables' variances and covariances are consequences of causal actions, and we aspire to understand observed variances and covariances by locating the underlying causal structures. Observed covariances or correlations do not come from the math or statistics of equations; they come from the causal world that underwrites the equations. The causal world also coordinates the means of the variables (Equation 2). If the causal variable takes on a value corresponding to its mean ($\bar{X}$), the resultant effect takes on a mean value ($\bar{Y}$).

If Y has two correlated causes, as in Figure 1B, the relevant equation is

$$Y = a + b_1X_1 + b_2X_2 \tag{5}$$

and the causal world demands

$$\bar{Y} = a + b_1\bar{X}_1 + b_2\bar{X}_2 \tag{6}$$

$$Var(Y) = b_1^2\,Var(X_1) + b_2^2\,Var(X_2) + 2\,b_1b_2Cov(X_1X_2) \tag{7}$$

The partitioning of the causal world partitions the variance in the effect (Y) but with the wrinkle that a portion of Y's variance comes from coordination/covariance between the values of the causes, not merely from variations in the values of those causes. This variance equation is fundamental to: understanding why some explained variance cannot be uniquely attached to a specific cause, understanding how biased estimates result from omitting correlated causes, and understanding what goes awry if an error variable covaries with a cause (for example, if X_2 was called a disturbance or error variable because it was not observed).

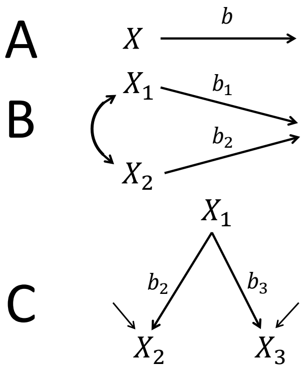

***Figure 1.* Basic causal structures.**

Figure 1C introduces X_1 as a common cause of X_2 and X_3, where the causal equations are

$$X_2 = a_2 + b_2 X_1 + e_2 \tag{8}$$

$$X_3 = a_3 + b_3 X_1 + e_3 \tag{9}$$

With independent errors/disturbances, this causal structure has the unavoidable consequence that

$$Cov(X_2 X_3) = b_2 b_3 \, Var(X_1) \tag{10}$$

The X_2 and X_3 covariance is called *spurious* because no direct effect links X_2 and X_3, but it is incorrect to describe the spurious correlation/covariance as "spurious (noncausal) associations" (p. 141). The association/covariance/correlation between X_2 and X_3 is the unavoidable consequence of the causal actions of the common cause, even if the relevant causal foundation is neither X_2 nor X_3 directly causing the other. The covariance in Equation 10 introduces the possibility of testing model implications, because estimates of the three right-hand terms can be obtained from the two model equations, and from data, without using $Cov(X_2 X_3)$. Comparing the model-implied covariance (from estimates of the two effects and the common cause's variance) with the corresponding observed covariance between X_2 and X_3 might (or might not) challenge the depicted causal structure. The failure of specific model-demanded causal consequences to match with observed covariances underpins model testing and diagnostics striving to improve models' causal structures.

The consequences of causal equations as presented above function the same way, whether the variables are observed or latent. And the consequences remain, even if some of the causally connected variables are observed while others are latent—which provides the principles grounding observed variables as measures of latent variables. Kline first introduces measurement of latents in Chapter 13 and in the context of factor analysis, where his emphasis on factors and outdated factor "rules" obscures the causal foundations of measurement, though measurement could have been helpfully introduced much earlier.

Equations 2 and 6 report that a case having a mean value on the applicable cause(s) is bestowed a mean value for the effect variable. Kline loses this easy and intuitive causal understanding when he turns to modeling means in Chapter 15, because he begins his discussion with non-causal regression. And he extends the confusion by referring to the a coefficients in these equations as "effects" of variables that are not variables "in the usual sense" (p. 371). That is, Kline omits cause from where it would be helpfully obvious, and adds cause where it really does not belong. Kline's new Chapter 8 demonstrates an emerging acknowledgment of the relevance, utility, and unavoidability of causal understanding of structural equation models, but he has not yet incorporated that understanding consistently throughout his text. Beginning with a clear causal emphasis would strengthen the text's foundational logic and encourage a focus on *principles* rather than *rules*. It would also reorient Kline's discussion of model testing and diagnostics toward checking and improving the model's postulated causal structures.

Readers seeking further instruction on the fundamental causal implications above, and unwilling to wait for Kline's fifth edition, might see Hayduk (1987) Chapters 1 and 2, and Hayduk (1996) Chapters 1 and 2. All the variance and covariance equations for models structured as in Figure 1 can also be derived as special cases covered by the matrix Equation 4.30 in Hayduk (1987).

Connecting the above to Kline's text

Kline's discussion of regression (Chapter 2) could have, and should have, differentiated between equations attempting to correctly represent a causal world and regression equations formed without requiring causal correspondence. SEM's concern for proper causal specification is fundamental, and hence the associated pedagogical issues dig deep. Kline's Equation 1.2 (p. 13), for example, makes it seem like covariances are just statistical rearrangements of correlations, rather than being the consequences of causal forces as in Equation 10 above. Covariance does not come from correlation, as this equation seems to imply. Both the covariance and correlation are consequences of some underlying causal world, and the SEM researcher's task is to ferret out the nature of that underlying world. Understanding how underlying causal structures produce patterns in covariances is what makes it possible to "understand patterns of covariances" (p. 14). Similarly, Kline's Equation 2.2 (p. 26) seems to say that the structural coefficient on the left of the equation somehow comes from the correlation and other terms on the right, when the structural effect would in fact produce, and be the source of, the correlation. Kline's Equation 2.2 is not wrong, in the sense that the entities on the two sides of the equation really are equal, but the arrangement of the equation and its surrounding discussion obfuscate how causal action produces the correlation. Kline's Equation 2.3 is a rearrangement of Equation 2 for means above, but Kline's explanation—which is essentially an assertion that the equation holds, and his calling this a "mean structure" (p. 27)—provides no hint of how causal action links the variables' means, and similarly fails to ground the reader's understanding in the easy-intuition that a case having an average value on the cause should have an average value on the effect.

And consider whether a goal of structural equation modeling is to "explain as much…variance as possible" (p. 14), or whether the goal is to accurately determine how underlying worldly causal forces produce and hence explain variances, covariances, and means. Regression can be sold as attempting to explain as much variance as possible, but it would be preferable to present structural equations as focusing on how variance and covariance are explained. Focusing on how variance is explained would clarify that there are wrong ways and right ways of explaining variance. This, in turn, focuses attention on the correctness of the model's specification, and clearly differentiates SEM from regression by revealing *how* regression equations can be wrong as causal equations. This would make it possible to avoid multiple awkward transitions between what are supposedly "regression" equations and associated wordings that are expressly causal. Readers wishing to understand and monitor the multiple diverse consequences of Kline's failure to ground his text in a search for worldly causal structures should consider Supplement Section 1. It is nice that readers report learning "something new" (p. 25) about regression from Kline's early material, but I would view it as more complimentary if readers had reported learning to differentiate between regression equations and structural equations pursuing causal understandings.

Time sequence and causal action

In Chapters 6 and 7, Kline is inconsistent in his consideration of time and causal action. He claims "presumed causes *must* occur before presumed effects" (p. 123; emphasis added, and see p. 296, 432, 465). He also says, "the absence of temporal precedence may not always be a liability when estimating reciprocal causation" (p. 137) and includes examples of reciprocal effects (p. 135, 136, 143, 151, 152, 154, 156, 186). Causal actions are more easily recognized and estimated when the cause occurs first, but it remains possible to estimate reciprocal and looped causal effects (Rigdon 1995).

Kline's time ambivalence can also be seen in the awkwardness of his attempt to differentiate between mediation and indirect effects (p. 134), but it approximates contradiction when he claims that non-experimental designs "cannot establish which of two variables, a presumed cause and a presumed effect, occurred first" (p. 124–25). Valid reciprocal-effect estimates can be obtained *without* determining which occurred first (Rigdon 1995), and this undercuts Kline's idea that reciprocal effect estimates in non-experimental designs are "in some sense" "always wrong" (p. 137).

Indeed, SEM developed in the social sciences because SEM made it possible to investigate causal actions in non-experimental designs without confronting the ethical difficulties accompanying social experiments. Rather than claiming that effects operating at congruent times are inappropriate or impossible, Kline should have instructed his readers on model features that make it possible to estimate, and check, reciprocal or looped effects (Hayduk 1987, 1996; Rigdon 1995); he might even have discussed how causal loops provide reinterpretations of models previously estimated without loops (Hayduk 1996). Estimated models in which a variable directly causes itself (e.g., Hayduk 1985, 1996) demolish the supposed requirement of temporal precedence, because a variable can't precede itself! Count me *out* of the supposedly "emerging consensus that mediation analysis requires data from designs with time precedence" (p. 141; and see p. 465).

A related concern arises in when Kline argues that reciprocal causal connections between variables makes it "plausible that they may share unmeasured causes" (p. 138), and hence that the reciprocally connected variables should be assigned covarying disturbances. If Y_1 causes Y_2 with no disturbance covariance, there seems to be no general reason that adding a reciprocal effect from Y_2 to Y_1 should automatically manufacture the existence of a "new" common cause producing covariance between the variables' disturbances. New causal actions from a common cause (namely, a kind of causal structuring requiring a disturbance covariance) do not pop into existence merely because of the existence of some other causal action (even if that other causal action forms a loop or reciprocal effect). Similarly the "bows" in Figures 7.1a, 7.2, and 10.7, and on page 143, are *not* necessitated by the reciprocal or loop causes in the figures, and the disturbance covariance reported in the last line of Table 14.6 (p. 351) lacks justification.

Observed variable and latent variable models

Chapters 6 and 7 consider the specification and identification of observed variable models, while Chapters 9 and 10 consider models containing latent variables. (The intervening Chapter 8 is addressed below.) These chapters are structured differently than in Kline's prior edition, and are afflicted by problems originating in: (a) the attempt to separate measurement from latent-to-latent effects (which obscures the advantages provided by modeling latent and observed variables simultaneously); and (b) the outdated presumption that latent variables require multiple indicators (p. 93). The chapter separation reflects the historical divide between path models and factor models, and seems to divide the rules for model identification into manageable chunks, applicable to first one part and then the other part of a model (p. 217). Unfortunately for Kline, the key advance provided by structural equation modelling was that it *overcame* the historical separation of measurement (via factor modelling) from structure (via path modelling) by combining and integrating measurement and structure. Some "strange" consequences of Kline's backsliding are presented in Supplement Section 3.

Kline's rules for separately identifying path-like and factor-like model segments are insufficient for full structural equation models, and new "rules" will be required for overall model identification. Model identification rules have lagged behind the melding of factor and path model components, and lagged even further behind for models containing: fixed coefficients (e.g., fixed measurement error variances), constraints between coefficients, causal loops (whether longer loops or self-loops), means, intercepts, moderators, multi-level components, and latent variables having no direct indicators. It is reasonable to attempt to ensure coefficient identification, but Kline seems to employ his rules as instructions limiting how to build models (p. 119), rather than granting researchers' theory and hunches primary control.

Kline displays considerable inconsistency in how he specifies and applies his rules. Sometimes "a single indicator is preferred" (p. 217); meanwhile, his identification Rule 9.1 (p. 201) requires three indicators for a single latent factor, or two or more indicators for each of two or more correlated factors. Thus, three indicators are identified, two indicators can be identified, and a single indicator may be preferred, yet we also read: "A better practical minimum is three to five indicators for each anticipated factor" (p. 195); "multiple-indicator measurement…is a cardinal characteristic of latent-variable models" (p. 127); and "each factor should have at least three indicators" (p. 454). If three indicators were really required, many of the models in Kline's Chapter 10 would be underidentified, because they contain latents having only two indicators—but they actually are identified, even though this is not evident because none of these two-indicator models were estimated.

Kline slants his advice to favour multiple indicators, but let's consider the second part of Kline's Rule 9.1 (p. 201)—namely, that a model would be identified with two or more correlated factors having two (or possibly more) indicators each. This identification "rule" really is *not* a general SEM rule, because SEM latent variables need not be "factors." Kline connects his Rule 9.1 to CFA (confirmatory factor analysis), but how is a reader new to SEM supposed to understand that latents need not be factors, and that factor models lack the latent level effects and constraints that potentially make measurement errors on even single-indicated latent variables identified? Figure 2 illustrates how latent causal connections can assist measurement identification in much the same way as do additional indicators. The measurement error variance for a single indicator (like X_1 in Figure 2) is often underidentified (unless provided a fixed value), but may be identified if the measured latent variable causes two or more latent variables, like η_1 and η_2. If the two causally downstream latents in Figure 2 are well identified and do not influence one another, this causal structure mirrors the three-indicator identification condition reported in the first part of Kline's Rule 9.1. Focusing on the consequences of causal actions makes it easier to appreciate the parallel between downstream-latents and downstream-indicators, while avoiding causal action and emphasizing the difference between latent and observed variables obscures the parallel. Identification of SEM measurements is not just a matter of a latent and its direct indicators. Measurement identification relates to how the latent fits into a causal network composed of both its indicator(s) and other latents.

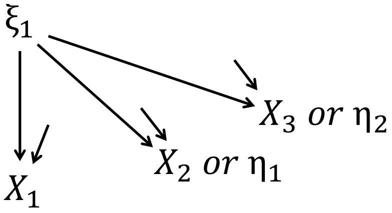

***Figure 2.* Latent effects potentially identify a single indicator's measurement error variance.**

Failure to appreciate how latent level structure can assist in estimating and validating measurement persists into Chapter 16 on measurement invariance and multiple-samples, and robs Kline of an opportunity to free SEM from some of its historical factor analytic entanglements. Kline understands that structural equation modelling is moving away from traditional EFA and CFA (see his identification "Rules for Nonstandard CFA Models," p. 202–06), but he seems not to recognize how the many other kinds of identification complexities render oblique rotations (p. 193), and rotational indeterminacy (p. 192) trivial tangents.

Concern for identification is relevant throughout the modelling process, not as in Kline's Figure 6.1 where identification is supposedly determinable *before* the selection of measures and remains unaltered by subsequent model revisions. Future texts would do better to present: the features that assist identification, the features that complicate identification, program output likely to appear for underidentified models, and ways of improving model identification (primarily introducing additional model or data constraints). I encourage readers to seek the structural model features underlying Kline's "rules" rather than memorizing the rules. Indeed, even models satisfying all the available identification rules can end up empirically underidentified due to random data variations or extreme values (p. 157, 206, 463), so even avid rule-followers should learn to recognize and respond to model underidentification.

A useful additional discussion of identification would address instances where the worldly model is underidentified given the researcher's limited data. Important disciplinary issues arise if the researcher's model is identified but the underlying worldly model is not identified given the available indicators. The more complex the causal world, the more likely it is that even nearly properly causally specified models will be underidentified, unless the researcher proactively addresses identification by employing causal variables entering the model at clear/precise locations, and constraining coefficients based on methodology or established "facts."

Chapters 9 and 10 introduce latent variables—first as factor-based measurements and later as effects between latents. Kline would have served his readers better had he begun by considering measurement of latents without factors. This would have forced a consideration of the difference between a factor and a latent variable. A *latent variable* is a variable or characteristic whose true values are presumed to exist and hopefully impact some indicator, though other "error" variables' causal actions prevent the latent's values from fully determining the indicator's values. Kline presents latent variables (Chapter 9, and p. 12–13) as if they are characteristics requiring multiple indicators—namely, as if they correspond to *factors*. Multiple indicators for a latent variable may be possible but are not required. Beginning with a latent variable like "age"—where true age differs from reported age (due to year-end jumps in reported age, memory, and avoidance of the next decade) would clarify that latent variables can have single indicators, and that researchers should address measurement error even with single indicators. Kline is simply wrong when he claims "multiple-indicator measurement…is a cardinal characteristic of latent-variable models" (p. 127; see also p. 213 last line, p. 220 second-last paragraph, and p. 223 third-last line). It is *not* the multiplicity of indicators that provides for latent variables. Latents are grounded in the acknowledgment of, and adjustment for, measurement error—where acknowledgment of, and adjustment for, measurement error can and should be done even with single indicators (Hayduk 1987, 1996; Hayduk and Littvay 2012). Latent variables are not necessarily factors, though *factors with multiple indicators* remain one style of latent variable. Second-order factors are latents having no direct indicators at all, and models may contain non-factor latents having no direct indicators (Hayduk 1990, 1996: Chapter 3).

Kline further confuses measurement when he equates latents with *constructs* and then says that constructs have different *facets*, as if latent variables also have facets (p. 127). A latent variable is a single, skinny dimension or number-line, and pretending that a single dimension has facets is nonsense. Speaking of facets and constructs demands causal connections between different latents, not reference to a single unidimensional latent.

Kline raises a different style of concern when he says, "Exogenous variable X_1 is assumed to be measured without error, an assumption usually violated in practice" (p. 213; and see similar statements p. 132, 352). This statement acknowledges measurement error but sounds as if the common failure to appropriately adjust for measurement error can be excused merely because many people have done this! By the next page we read, "Because is specified as exogenous, it is assumed to have no measurement error" (p. 214). Unfortunately, being exogenous does not alter the existence of measurement error, and hence Kline has turned the common violation from the previous page into an "assumption," as if this does not become a "usually violated" and deficient assumption. There simply is no justification for failure to compensate for a reasonable amount of measurement error in exogenous variables.

Measurement error in endogenous latent variables is equally important, though also miscommunicated by Kline. In reference to Figure 10.1a, Kline makes a statement that is appropriate for some endogenous variables (namely, those having no further downstream latent-level effects) as though it applies to all endogenous variables. "This assumption {namely of no measurement error} is not required for the endogenous variables in this model, but random error in Y_1 or Y_3 is manifested in their disturbances" (p. 213; { } material inserted). This statement is arguably true for Y_3 in the relevant model, but it is definitely false for Y_1. For endogenous latents having no further latent effects (like Y_3) the only consequence of measurement error variance is to increase the variance of the "disturbance" variable attached to the modelled endogenous variable (which makes the disturbance a mixture of measurement error and other omitted causes of the latent). The statement is false for Y_1 in the Figure 10.1a model, because this variable's measurement error functions *causally differently* (has different implications) than real omitted causes of the latent variable. Measurement error for a latent in the diagrammed Y_1 position would causally impact only latent-Y_1's indicator, while disturbance-style omitted causes of latent-Y_1 would impact both latent-Y_1's indicator and the causally downstream variable Y_3. The differing implications introduce model misspecification, unless the model differentiates between measurement error and latent-level structural disturbance. (Kline makes the same mistake regarding X_1 in Figure 10.2a, and Y_1 in Figure 10.2b.) Rather than attempting to specify a new rule indicating when a researcher must adjust for measurement error in endogenous latent variables, it is simply safest to routinely acknowledge and adjust for measurement error in all latent variables all the time (Hayduk and Littvay 2012). And the adjustments should strive to attain validity, not merely "control for score reliability" (p. 127).

Contrary to Kline's claim that "there is no point in retaining a model with just as many explanatory entities (factors) as there are entities to be explained (indicators)" (p. 190), there may be a very important point to be made. Differentiating causal actions dead-ending in each indicator (namely, measurement errors) from causal actions impacting the latent's true values, which subsequently cascade to downstream latent variables, might require a model containing as many latents as indicators. The latents having single indicators would not be "factors" but SE models need not contain "factors."

The diagrammatic partitioning of variance in Figures 6.3 and 9.2 (p. 131 and 190) differentiates between measurement error and real variations having unknown causes, but Kline pays insufficient attention to properly causally modelling the various variance components. For example, Kline recommends using scale reliabilities to determine fixed-measurement error variances (p. 223; and see p. 458). This adjusts for some measurement concerns, but is insufficient because proper modelling requires appropriate causal differentiation between all the indicators' components (not just a reliability adjustment), and requires attention to their covariances, not just their variances. In the context of scales, the concern for validity recommends modelling all a scale's items, not just scale scores.

Factors versus factor scores, latents versus latents' values

A different but related issue is whether SEM or SEM-measurements require that we know the values of latent variables. We sometimes would like to know the values of latents—for example, if the latent is someone's ability and a related decision is required—but are we *required* to know the true values for latents in SE models? Think again of age. Are we required to determine the cases' true ages in order to use latent age in a structural equation model? The answer is definitely no, we do not require the latent variables' values. The complement of this is that we equally definitely do not need to know the values/magnitudes of error variables differentiating true latent variables' values from observed indicators' values. The statistical "magic" grounding the ability to estimate latent-level coefficients without knowing latent variables' values can be seen in moving from Equation 4.28 to 4.29 in Hayduk (1987), but is a bit too cumbersome to detail here. The relevance of this for Kline's text comes from *factor score* attempts to estimate the values of factor-latent variables. Factor score *indeterminacy* (p. 189, 193, 212) refers to the indeterminacy or imprecision in possible sets of estimated latent-factor values derived from the observed indictors. Kline says that "because theoretical variables and their proxies (indicators) are almost never identical, estimates of causal relations between latent variables are approximate at best" (p. 212). Let us repeat Kline's statement, replacing "theoretical variables" with "latent variables" and "proxies (indicators)" with "factor scores," estimated from the indicators, as Kline's context requires. "Because *latent* variables and their *factor scores* are almost never identical, estimates of causal relations between latent variables are approximate at best." To be even clearer, *because* we can't determine latent true scores precisely, estimates of the effects between latent variables are supposedly approximate at best.

Unfortunately, Kline's "because" is entirely unfounded. Obtaining proper effect estimates does not involve, require, or depend on latent variables' true scores. Contrast Kline's claim that latent-to-latent effects are "approximate at best" with the observation that latent effect estimates can be more accurate and precise because they adjust or compensate for measurement error, while the corresponding estimates from observed variables are prone to contamination with measurement error. (This was historically called *correcting for attenuation*.) Kline employs something that is not required in SEM (factor/latent scores) to inappropriately reframe a strength of SEM (adjustment for measurement error) into a supposed weakness!

Factor-score indeterminacy also seems to underlie Kline's warnings against "jingle-jangle fallacies" (p. 301, 458), where a single name is insufficient to force indicators to reflect a single latent (jingle), or different names are insufficient to force indicators to reflect different latents (jangle). More cogent warnings could have been made in the context of descriptions or interpretations of latents as encapsulating verbal understandings/theories, but where those understandings/theories may be inconsistent with the causal world controlling the indicators.

I end this section by noticing that factor-structured sets of indicators tend to fail because the required common-cause structuring is often inconsistent with the causal forces actually producing the indicator data. Kline's only real example of factor analysis appears at the end of Chapter 9 (p. 206–08) and is highly significantly inconsistent with the data. The inconsistency is not reported in the text, though it can be found in the publisher's model archive. Some sleight-of-the-writing-hand switched the strong *dis*-Confirmation of this factor model into a *Con*firmation, permitting the section to be titled a "CFA Research Example" (p. 206). Kline says that "relatively few applications of CFA are strictly confirmatory" because post-hoc model modifications are introduced to make the model fit, and/or because CFA is claimed after exploring with EFA, but it seems that even clear and direct model failure also fails to tarnish the *C* in CFA.

Chapter 8, Structural causal models

It is a breath of fresh air to encounter Chapter 8's focus on causal actions and the unavoidable implications of causal actions. This chapter is new to Kline's fourth edition, presenting jargon introduced by Judea Pearl and "rules" reporting the unavoidable implications of directed acyclic graphs (DAGs). Kline makes a truly admirable attempt to present this material, but it is sufficiently complex that I offer some nit-picking suggestions (in Supplement Section 4), though I have two more substantial concerns. The first is that Kline does not introduce Pearl's *do*(*x*) operator (Pearl 2000: 70), which is fundamental to understanding the propagation of causal effects. The *do*(*x*) operator permits following the unavoidable consequences of *do*ing or making specific, precisely expressed hypothetical interventions within the modeled causal structures. By excluding this, Kline loses an opportunity to assist his readers to see SEM's fundamental causal precision in action. The second concern is that Kline seems to have missed the consequences of Chapter 8 for his other chapters.

This causally focused chapter contrasts with Kline's repeated avoidance of cause, as discussed above. The precise and unavoidable causal implications discussed in Chapter 8 should have fortified Kline's Chapters 11and 12 on model testing. Kline has argued against coefficient testing in other contexts (Kline 2013), but Chapter 8 provided a missed opportunity to clarify the fundamental difference between coefficient tests and model tests. DAG tests are tests of precisely demanded model/theory implications—namely, they are tests that incorporate and investigate the combined consequences of multiple theorized causal structures, and differ from direct tests of specific coefficients.

Chapter 11, Estimation and local fit testing

Chapter 11 reads smoothly but is sprinkled with multiple problematic statements. One easily misconstrued claim is that single-equation estimation methods "can be less affected by specification error than simultaneous methods" (p. 231; and see p. 235). The misconstrual comes from failing to consider the other, not-less-affected equations, and from the single equation methods discussed being routinely misspecified because they fail to compensate for measurement error (see p. 233, point 1). There are single-equation estimation methods that address measurement (e.g., Bollen et al. 2014) but these are not discussed, and have their own limitations. Another misconstrual is embedded in Kline's description of maximum likelihood estimates as "the set of parameters estimates that is most likely to have generated the observed data" (p. 236). The misconstrual should become clear if you hear this an an example of "the false belief that *p* measures the likelihood that H_0 is true, given the data" (Kline 2013: 98).

But a more fundamental concern is Kline's inappropriate adoption of a specific correlation-discrepancy size of .10 as his implicit *local fit test* criteria. Smaller discrepancies between observed and model-implied correlations are presumably acceptable, while discrepancies "of .10 or more may signal appreciable model-data disagreement" (p. 240). Kline repeatedly appeals to whether or not .10 is exceeded (p. 253–54, 278, 329, 380, 385, 408, 416, 481) and speaks as if this constitutes reasonable local fit testing (p. 241, 283). Unfortunately, exceeding or not-exceeding a correlation discrepancy of .10 does not constitute a statistical test, and the .10 value lacks statistical justification. The value merely demarcates the boundary between what Kline will attend to or disregard. (Actually, Kline does not even stick consistently to his .10 value (p. 344).) Programs like LISREL and EQS report standardized residuals (p. 252), which provide statistically appropriate local fit tests, but Kline displaces the available tests with a "criterion" more to his liking, even though he knows important model misspecifications can produce only smaller amounts of model-data discrepancy (he cites Hayduk 2014a), and even though he has heard that "Shame for disrespecting evidence {will constitute one of} the personal consequences of insufficient respect for structural equation model testing" (p. 496; {} material added).

Combining "Estimation and local fit testing" into a single chapter has the unfortunate consequence of placing local fit testing prior to overall model testing (namely, Kline's next chapter). Both model testing and local fit testing depend on estimation, but overall model testing should precede investigation of local fit. For a bad example, notice that after dedicating more than a dozen pages to a detailed example, Kline says "the fit of the example model is unacceptable" (p. 253) both locally and globally, but he fails to respond appropriately to the model-data inconsistency. Changing the model's structure to conform to the world's structure would: alter the control variables, alter the estimates, alter the estimate's significance, alter the basis sets, alter the residual ill fit, and mess with just about every claim Kline made about his example. Kline's comment regarding model-data inconsistency does not even hint at the numerous diagnostic investigations that should be undertaken, or the substantial model reassessments that should accompany detection of model-data inconsistency. It is easy—deceptively easy—to think that if a model "poorly explains certain observed associations" (p. 253), the model problems are tightly linked to those particular problematic covariances/correlations rather than being detectable symptoms of more dispersed yet fundamental model misrepresentations. Localized ill fit does not confidently report localized model specification problems. Patterns of local ill fit can sometimes contribute usefully to diagnostic examinations, but even patterns of local ill fit do not irrefutably detect the relevant model problems. I count it as a serious deficiency that the term "diagnostics" does not even appear in Kline's subject index. Diagnostics require assessing the many things potentially wrong with a model, and/or the data, rather than routinely freeing the nearest error covariance—which often amounts to blaming/convicting the nearest bystander.

Chapter 12, Global fit testing

Chapter 12 also reads smoothly, but it is perhaps Kline's most problematic chapter. The problems begin with the title. Kline knows "that there is actually no dependable or trustworthy connection between the size of the residuals and the type or degree of model misspecification" (p. 278)—where residuals refer to the difference between the model-implied covariances/correlations and the data covariances/correlations. Given the *disconnect* between the amount of ill fit and the seriousness of model misspecification, researchers face a choice of being interested primarily in model misspecification or model fit. Even a brief consideration determines the primary concern is model misspecification, while fit plays only a supporting role. Researchers want to test their *model*, not just their model's fit, and examine fit to see whether or not this provides evidence of model misspecification. Thus, a more appropriate chapter title would be "Testing for model misspecification" or "Detecting model misspecification" rather than the current "Global fit testing."

If ill fit is detected, researchers should probe the program output for diagnostic clues to the nature of the problem(s) and potential model or data improvements. A ringing ill-fit alarm bell should prompt thorough and detailed investigation of possible data mistakes as well as the multiple kinds of possible model misspecifications—not mere pursuit of different fit-index ways of reporting the ill fit (p. 266).

Kline grounds his fit-index-based disrespect for test evidence in an oft-cited claim by Box (1976: 792) that "all models are wrong" (p. 263)—which implicitly and inappropriately suggests that model misspecification cannot be avoided, so you shouldn't worry if you encounter some. God might know whether or not all models are wrong, but how could even a famous person like Box know about all structural equation models—including models that have not yet been specified or run? In fact, Box was not even referring to structural equation models—he was writing back in 1976, when structural equation modelling was relatively unknown, and SE model testing nearly nonexistent. Incompleteness can make many styles of statistical models wrong, but incompleteness does not necessarily contribute to SEM ill fit, because measurement errors and unknown latent disturbances are *parts* of structural equation models, not omitted features. And given that we can construct structural equation models of experiments, applying Box's statement to SEM implicitly asserts that all experiments are wrong—because SE modeling of experiments would also only result in wrong models! Claiming that all experiments are wrong is clearly nonsense.

None of these kinds of considerations have stopped Kline (and some others) from propagating this nonsense and its paraphrases. See: "When (not if)" the model does not fit (p. 120), models "are imperfect approximations" (p. 262), and fit indices "allow for an 'acceptable' amount of departure from **exact (perfect) fit**" (p. 60; emphasis in the original). Correctly specified models may be rare (p. 232), but rare does not mean impossible (for fitting models, see Entwisle et al. 1982; Hayduk 1994; Hayduk et al. 1997, 2005).

Kline ends his quote from Box (p. 263) a bit too soon. Box's next sentence reads: "Since all models are wrong, the scientist must be **alert to what is importantly wrong**" (Box 1976: 792; emphasis added). Kline stumbles regarding what is "importantly wrong" because he frequently conflates importance with the amount of ill fit, rather than with the nature of underlying model misspecifications. For example, Kline provides a section on a "Recommended Approach to *Fit* Evaluation" (p. 268; emphasis added) rather than to *Model* evaluation. By focusing on *Approximate Fit Indexes* (p. 266–68), RMSEA (p. 273–75), CFI (p. 276), and SRMR (p. 277), and continuing with these indices in his examples, Kline distracts from a search for "what is importantly wrong." Kline notes that there are "discredited thresholds for such fit statistics" (p. 269, 268), without reporting which specific thresholds have been discredited, and despite his continuing use of thresholds courting discreditation (p. 267, 274, 276–78).

Kline fails to appreciate the depth of the challenge to ALL model fit indices created by there being "no dependable or trustworthy connection between the size of the residuals and the type or degree of model misspecification" (p. 278). He proceeds as if small-sized residuals overrule or overturn the significance of those residuals, whether in the context of global fit testing or in local fit not-real-testing, via his indefensible .10 correlation residual (p. 462, 265). This can be seen in Kline's claim that a large *N* devalues χ^2 testing because a larger *N* enables χ^2 to detect smaller covariance/correlation residuals—including residuals smaller than his .10. For many *misspecified* models, χ^2 power does increase with *N* and thereby demonstrates increasing power to detect misspecifications, but for properly specified models, χ^2 does *not* increase with increasing *N* (Hayduk 2014b). Kline inappropriately reports χ^2 as being "overly sensitive to sample size" (p. 271; see also p. 330, 462), when in fact χ^2 increases with *N* only when there is some detectable problem in the model or data. Kline's Exercise 4 (p. 279, 298) and its suggested answer (p. 480) are ill-founded, because with a proper model and larger *N*, the data covariances would become more stable due to smaller sampling variations from the true covariances, and consequently χ^2 would not inflate. The idea that χ^2 is "overly sensitive" implicitly appeals to there being discrepancies that are too small to be worth detecting and investigating, when in fact covariance discrepancies can be zero even in the presence of important model misspecifications (Hayduk 2014a). Kline's proclivity to think of large-*N* χ^2 as detecting trivial fit differences parallels his tendency to disregard local ill fit correlations less than .10, even though there is no justification for claiming that smaller residuals in either context protect researchers from important model misspecifications.

A related imprecision is Kline's failure to distinguish between the causal structure of a model and the fit provided by that model. For example, Kline titles one section "Equivalent CFA models" (p. 315) and another "Equivalent SR models" (meaning *Structural Regression models*; p. 348), when what he is discussing are causally *non*-equivalent models providing equivalent fit. The models are not causally equivalent because they contain different causal specifications, though they have corresponding covariance implications. Had Kline's titles been something like "Different CFA or SR models producing equivalent fit," the discussion would have turned to causal specification/misspecification, and the models reported on pages 345 and 358 would have been described as *not distinguishable on the basis of their covariance fit*, even though the worldly causal structures are empirically distinguishable.

Kline's discussion of RMSEA (Root Mean Square Error of Approximation) contains multiple technically correct statements but is unlikely to assist anyone not already familiar with the RMSEA (p. 273), and actually encourages problematic SEM practice. Kline cites a 2008 work as indicating there is "little support for a universal threshold of .05 (or any other value)" for the RMSEA (p. 274), yet spends the next pages propagating obsolete threshold values suggested by Browne and Cudeck back in 1993, only to follow this with additional relatively recent references that "question the generality of thresholds for the RMSEA" (p. 276). As if these were not enough, Kline disregards the logical problem at the heart of the RMSEA.

The problem is that non-zero RMSEA criteria attempt to excuse or overlook some amount of real model-data inconsistency for each model degree of freedom. The RMSEA is calculated as *ill fit per degree of freedom*, and hence claiming a non-zero RMSEA value as acceptable claims that some non-zero amount of real ill fit is acceptable for each and every model degree of freedom. (The Browne and Cudeck reference Kline cites as foundational for the RMSEA describes real (non-random) model-data inconsistency as "error of approximation" (1993: 141).) Overlooking, excusing, or discounting real model-data inconsistency is clearly problematic. And a model rendered strongly testable by having many degrees of freedom is supposedly excused (ahem) from that strong-testing because Browne and Cudeck said that the RMSEA permits (cough) overlooking some amount of model-data inconsistency for each degree of freedom—an amount not based on statistics but "based on subjective judgment" (1993: 144).

Kline (p. 274) cites work documenting a clear instance where the same Michael Browne (of Browne and Cudeck), along with Robert MacCallum (another big name who championed disregard of real model-data inconsistency) and Kim, Andersen, and Glaser (2002), defended and retained a model that was inconsistent with their data. Browne et al.'s "supposedly negligible ill fit obscured important, systematic, and substantial causal misspecifications" that were located and corrected by attending to relevant experimental conditions (Hayduk et al. 2005: 1). Somehow Kline remains immune to the methodological unacceptability of overlooking real model-data inconsistency, despite: the "critical" RMSEA value being subjective (not statistically based), challenged by multiple recent references, and having led strong people into making indefensible modelling mistakes.

Unfortunately, Kline perpetuates this problem throughout the remainder of his book. It is nice that Kline's testing chapter summary reports there being a "consensus that some routine practices are inadequate" (p. 297), and nice that his best practices chapter says, "Do not rely on 'golden rules' for approximate fit indexes to justify the retention of the model" (p. 461), but these come across as disingenuous, given that he also says, "If possible, report…the RMSEA" (p. 464), given that the logical problems with indices are not respected in Chapter 12, and given that the RMSEA is repeatedly incorporated later (p. 279, 290, 305–07, 317, 328, 344, 347, 350, 357, 359, 380, 385, 406, 416). The same inconsistency appears when Kline complains that low power provides "little chance of detecting a false model" (p. 265), while encouraging low power—by appealing to fit indices rather than tests, by denigrating the power provided by larger *N*, and by switching from a test of model fit to the hypotheses of "close fit" and "not-close fit" (p. 290–91). If he really appreciated power, he would have considered power in the context of the most powerful available test—namely, the χ^2 model test—not merely any lower-power index. In fact, power is defined only in the context of model testing, not indexing, so a statistical sleight-of-hand accompanies discussing the power to detect arbitrarily hypothesized index values (p. 292, Figure 12.2 and fourth line). In short, Kline has not yet come around to consistently reporting the deficiencies of the RMSEA and warning against its continued use.

Unfortunately, Kline's text falters in several additional places due to deficient model testing, as you can see by considering the technical-teasers in Supplement Section 5. I assisted the third edition of Kline's *Principles and Practice of Structural Equation Modeling* by providing some "backbone" (Kline 2011: *xi*) to his testing chapter, only to find that his fourth edition reverted to the **deficient** view that there should be "LESS EMPHASIS ON SIGNIFICANCE TESTING" and that the "proper role for significance testing in SEM is *much* smaller" (p. 17; both emphases in the original). I hope the above constitutes a sufficiently clear and strong prosthetic to propel Kline, or his successors, into understanding the important difference between a model being significantly inconsistent with the data versus "acceptably close" to the data (p. 11).

Chapters 13 and 14

"Confirmatory factor" and "Structural regression" models are addressed in Chapters 13 and 14, respectively. Factor analytic measurement of latent variables developed quite independently of regression/path-analytic linking of different variables, and this led to a common presumption that structural equation modelling should proceed in two separate steps—initial measurement of latent variables via factor analysis, followed by structural connections between different latents. Kline cites Anderson and Gerbing (1988) on two-step modelling and aligns himself with the two-step approach by separating these chapters. This supposedly reduces the complexity of model assessment by providing a "separation of measurement issues from structural issues" (p. 340).

Unfortunately for Kline, one of SEM's greatest strengths is to combine, not separate measurement and structure. The reason is simple. Structural equation modeling strives for *valid* models and *valid* measurement, not merely reliable models and reliable measurement. Measurement validity requires that a measured *latent variable function appropriately* in connection to *other latent variables*. The possibility of demonstrating appropriate connections to diverse theory-specified variables only arrives with the latent level of the model, and hence measurement remains incomplete until the measures are incorporated into full, well-functioning models. All of a latent's indicators (not scales or parcels) should be retained in the full SE model if a latent proposed by factor analysis seeks validation.

Kline's emphasis on multiple factor-structured indicators results in insufficient attention to measurement error variance in routinely used variables like age, sex, and education, which rarely have more than a single indicator. Another consequence of requiring multiple indicators is that this tends to displace latents which could function as control variables, instrumental variables, or variables clarifying the operative causal mechanisms. Kline cites Hayduk (1996) but somehow misses its Chapter 2, which explicitly challenges Anderson and Gerbing's two-steps; he also cites Hayduk and Glaser (2000a), but somehow also misses the corresponding challenges to four-steps. Kline is seriously off-base when he cites Hayduk and Glaser (2000a, b) as supporting either two- or four-step modelling (p. 339). In short, by recommending use of "two-step modeling, not one-step modeling" (p. 462), Kline renders his book incapable of providing a thorough discussion of measurement validity, burdens his reader with identification "rules" that really aren't sufficient-rules, misses adjustment for measurement error in variables like age, and hinders pursuit of informative model specifications.

In Chapter 14 we re-encounter the problem of waffling between causal and non-causal connections between variables. A conscientious reader will cringe at the chapter title's reference to *Structural Regression Models*, because structural effects at the latent level are not mere regression coefficients. Kline's sporadic attention to latent-level causal structuring is evident in his "Detailed Example" (p. 341–48), which considers a model by Houghton and Jinkerson (2007). Kline says these "authors describe the theoretical rationale for each and every direct effect among the four factors in the structural model" (p. 220), which exudes careful causal attention. Then we notice that four of the six latent-level effects in the model Kline "would retain" (p. 347) are insignificant (Figure 14.2, p. 348). The insignificant effects disconnect the *Constructive Thinking* latent from the remainder of the model's latents, including the final dependent variable *Job Satisfaction*. The absence of evidence supporting these effects contrasts sharply with Houghton and Jinkerson's claim that their study results "suggest that *constructive thought* strategies are related to *job satisfaction*" (2007: 51; emphasis added). Surely either the theory or measurement are problematic if the measurements do not locate latent variables displaying causal actions required by theory! Several modelled indicators had been constructed by segmenting sets of items into parcels, or sub-scales, whose reliabilities were previously reported (see p. 220; and Houghton and Jinkerson 2007) and several of the items had uncomfortably low proportions of explained variance. But the causal disconnection seems not to have prodded any reconsideration of the measures. Somehow both the measurement and theory emerge unscathed, despite the vanishing theorized effects!

Further causal imprecision appears in this example when Kline adds a measurement error covariance between the indicators of Happy and Mood_2 because there is "common item content across the two indicators" (p. 344). Unfortunately, the estimated covariance is inconsistent with this common content theory, because it is negative (p. 347)! Common content would be consistent with a positive covariance, but not a negative, measurement error covariance. But why might Kline have missed this? Notice that Kline's justification appeals to "common item content," not to a common *cause* of the items. Had Kline thought of his justification as requiring a common cause, that should have triggered considerations paralleling Figure 1C and Equation 10 above, which report that a commmon cause only produces a negative covariance if the effects have opposite signs—but opposite signed effects are incompatible with Kline's appeal to common content. The unjustified negative coefficient seems to "'clean up' local fit problems" (p. 345), and provides a clear example of how causally inappropriate and likely misspecified coefficients can deceive researchers by sopping up local ill fit. Sure, the fit is improved, but the properness of the model's causal specification has been sacrificed to attain the improved fit.

Kline's example contains yet another missed opportunity. The measurement error covariance that Kline added to his measurement model results in his model having one fewer degree of freedom than Houghton and Jinkerson's corresponding model (Kline's 47 versus their 48). Their final model restricts two additional effects and leads to Kline's model having two more degrees of freedom (49); but Houghton and Jinkerson (2007: Table 2) report three more degrees of freedom (51 rather than 50) for their final model. Kline's Chapter 6 exercises instruct readers in counting degrees of freedom, but Kline misses this opportunity to apply this skill. One latent variable in Houghton and Jinkerson's final model seems to have been scaled in two different ways, namely, by setting a "loading" to 1.0 and by simultaneously setting the latent variable's variance to 1.0, like all their other latent variables (Houghton and Jinkerson 2007: 47, Figure 1). This double-scaling forces their *Job Satisfaction* latent to contribute exactly 1.0 unit to the variance of their *Job Satisfaction* indicator (like setting both the variance and effect in Equation 3 above to 1.0 if Y was the indicator), when that indicator actually has a variance of $0.939^2 = .88$. This forces a model-data inconsistency, explaining why Kline's model fits better than their model, and should have produced an impossible negative error variance estimate for Houghton and Jinkerson's scaling indicator. Kline missed this opportunity to show how model misspecification can lead to problematic estimates, and missed the opportunity to caution that the literature contains enough errors to warrant routine caution and checking.

Chapter 15

This chapter addresses modelling means and latent growth curves. It reads smoothly but requires revision because it both omits causation where it is appropriate, and inserts causation where it is inappropriate! Equations 2 and 6 above illustrate how intercepts coordinate the means of causally connected variables, and this parallels Kline's Equation 2.1 (p. 369 and 27), but Kline flounders because his equation is presented in the context of regression rather than causal action.

In the example spanning pages 369–72, Kline interprets the intercept of 20.000 in his regression equation

$$\hat{Y} = 20.000 + .455(X) \tag{11}$$

as "the direct effect of the constant on endogenous variable Y" (p. 371). The "constant" is a 1.0 value placed behind the intercept, which permits rewriting the equation as

$$\hat{Y} = 20.000(1.0) + .455(X) \tag{12}$$

and on mid-page 370 as

$$\overline{Y} = 20.000(1.0) + .455(\overline{X}) \tag{13}$$

Unfortunately, it is incorrect to interpret the intercept 20.000 as "*the direct effect of the constant* on endogenous variable *Y*" (p. 371). One way to see why begins with noticing that the intercept depends on the scales of *X* and *Y*. If the scale for Kline's *X* variable (Table 15.1, p. 370) had resulted in each case's *X* score being 10 units higher, the *X* mean would increase by 10, and *X*'s mean contribution in Equation 13 would increase by 4.55 units; the intercept would correspondingly decline by 4.55 units, from 20.000 to 15.45. This imagined change in *X*'s scale alters the intercept *without* changing *Y*, the 1.0 constant, or any variable's effectiveness—namely without changing *any* feature in Kline's interpretation of the intercept as "*the direct effect of the constant* on endogenous variable *Y*"! Similarly, had *Y*'s values been reported on a scale reading 20 units lower, Kline's equation would become $\hat{Y} = 0.0(1.0) + .455(X)$ and the "constant" would now seem to have no "effect" (according to Kline), even though this change in *Y*'s scale does not change the causal effectiveness of anything.

But deeper causal issues are also involved. Consider the word "constant" in describing an intercept as "the direct effect of the *constant* on endogenous variable *Y*". Can a *constant* have effects, direct or otherwise? The answer is no—constants do not have "effects," because effects demand potentially different outcomes resulting from *variations* in the cause (see Pearl's definition; 2000: 70), or Mulaik's insistence on variables (2009: 84–102)). Kline understands "the constant…is not a variable in the usual sense, because it has no variance" (p. 371), but he persists in referring to the intercept as the "effect" of the constant even though there is no such thing as an effect without potential variability in the cause. Placing a 1.0 after the intercept mimics the positioning of a causal variable like *X* but is insufficient to warrant interpreting the 20.000 intercept as an effect.

Next, consider that just as error variance changes upon addition of new predictors, intercepts also change upon addition of new predictors. And just as we do not know which real variables contribute error variations in equations, we do not know which real variables contribute to the intercepts in equations. Each equation's error variable's variance stands-in for the variation-consequences of unspecified causal variables, and each equation's intercept stands-in to coordinate the effects and scales of all the unspecified causal variables, with the effects, scales, and means of all the included variables. Failing to acknowledge intercepts' connections to real excluded variables leads Kline to the problematic claim that "If a variable is excluded from the mean structure, its mean is assumed to be zero" (p. 372). That is false because including a new clause adds a term containing the effectiveness and mean of that variable to the right of equations like Equation 13. A *regression* error variable is assumed to have a zero mean, but a real excluded causal variable need not have zero mean. If the *Y* variable in Equation 13 was exogenous because it had no specified causes, the *Y* mean would correspond exactly to the intercept but that would not force the unmodeled causes of *Y* to all have zero means.

Now consider that each different equation's intercept, and each different exogenous variable's mean, depends on a unique set of excluded variables having different scales, means, and degrees of effectiveness. Contrast this with **the** *constant* in Kline's intercept interpretation, represented by **the** triangle-containing 1 in Kline's Figure 15.1 (as duplicated at the top of Figure 3). Kline's figure represents **the** *constant* as the (singular) common cause of two variables, and as the common cause of many more variables in figures on pages 385, 389, 404, and 412. If the constant really was a common cause influencing two downstream variables, this would result in covariance between the two effected variables (recall Figure 1C and Equation 10 above). Unfortunately, Kline's representation is inaccurate: the arrows connected to the triangle-1 are not effects, the triangle-1 is not a variable, and the implication of covariance would be wrong. Each endogenous and exogenous variable should be granted its own "constant" corresponding to the unique set of unmodelled variables contributing to the constant following/attached-to the intercept for each endogenous variable, or to the mean of each exogenous variable, as depicted lower in Figure 3. There the dotted lines are obviously not effects, because they have no arrowheads, and each variable is granted its own constant—whether representing an intercept or exogenous variable's mean. This representation is similar to that employed by Hancock, Kuo, and Lawrence (2001), but the absence of arrowheads ensures these are not interpreted as "effects," and deletion of the triangled-1 minimizes space requirements. Placing the intercept/mean designation near the variable's unmodeled sources of variance (disturbance/error variance for endogenous variables; total variance for exogenous variables) signals that the unmodelled causal variables contributing variance also contribute to the corresponding variable's mean or intercept.

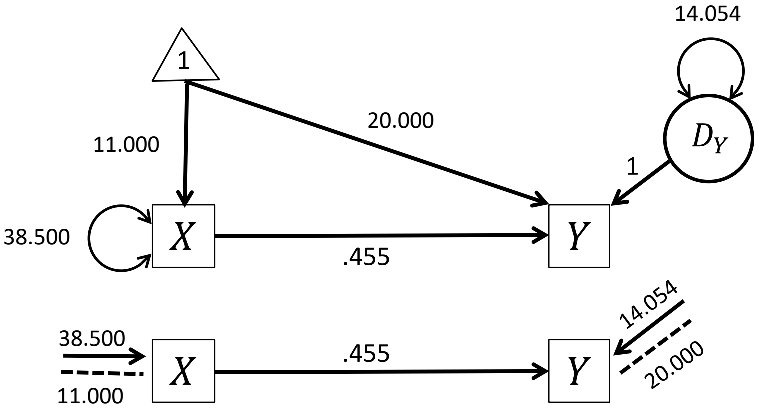

***Figure 3.* Kline's Figure 15.1 (top) and a replacement (bottom).**

The literature is not yet committed to a single felicitous mode of diagramming means and intercepts, but we can, and should, be consistent in our verbal descriptions. Consider part 2 of Kline's Rule 15.3, which claims that "for endogenous variables, the direct effect of **the** constant is an intercept but the **total effect** is a mean" (p. 372; emphasis added). The intent of this "rule" is reasonable, but its execution is problematic. The intention behind "total effect" can be seen by writing the equation for *Y*'s mean in Kline's Figure 15.1 as

$$\bar{Y} = 20.000 + .455(\bar{X}) \tag{14}$$

and rewriting this with the numerical *X* and *Y* means, and inserting the constant (1.0):

$$25.000 = 20.000(1.0) + .455(11.000)(1.0) \tag{15}$$

This equation corresponds to the math on pages 370 and 372, and makes it *look like* the *Y* mean (25.000) is the sum of a direct "effect" of the (1.0) working through the intercept of 20.000 and an "indirect effect" of the (1.0), calculated as the produce .455(11.000)—so Kline expresses the *Y* mean as a total effect of the constant. The problem is that the intercept (20.000) and *X* mean (11.000) are not effects, and there really is not *one* single/lone constant. Each (1.0) refers to a *different* set of variables. In Equation 15, the left-most (1.0) refers to the disturbance/error/unavailable causes of *Y*, while the right-most (1.0) refers to entirely unknown causes of the *X* mean. The (1.0)s look the same, and the triangle-enclosed 1 in Kline's Figure 15.1 looks like a single variable—but it isn't. There are in fact multiple (1.0)s, and each refers to a different set of variables.

This recommends modifying Kline's statement quoted above to say that an endogenous variable's mean can be calculated as the "total effect" of *multiple different sets of variables*—which is glaringly self-contradictory, because total effects tabulate effects originating from a single source, not effects originating from multiple different sources. Put simply, the verbal reference to "*the* constant" inappropriately conflates multiple different causal entities, and recommends multiple corrections to Chapter 15 (see p. 371, 372, 378, 379, 380, 383, 384, 386, 387, 388) and Chapter 16 (p. 403, 420). The calculations of product terms work largely as Kline reports, but Kline's linguistic and diagrammatic descriptions of why things work that way require substantial revision.

In addition to the above, Kline's presentation of means and intercepts would benefit from observing that many social science variables have arbitrary scale-origins and scale-units. This would provide an opportunity to address the difficulties involved in locating non-arbitrary scale origins and units of measurement, as well as clarify why even overidentified SE models with mean structures have a limited ability to assist in locating non-arbitrary means and intercepts.

Kline's detailed example of means and intercepts

Kline illustrates the modelling of means and intercepts using data from military personnel repeatedly attempting an air traffic controller exercise (p. 375–87). Kline bases his example on 137 cases that Browne and Du Toit (1991) selected for reanalysis from experiments conducted by Kanfer and Ackerman (1989). I suspect, but was unable to confirm, that the cases came from Kanfer and Ackerman's third experiment, and I was unable to determine how these cases had been selected from Kanfer and Ackerman's many cases, or even whether these cases came from an experimental or control condition. We cannot hold Kline responsible for Browne and Du Toit's failure to report how they selected their cases, but we should hold Kline responsible for emphasizing an example that precludes careful consideration of the data-gathering details that constitute the bedrock of competent structural equation modelling. As we shall see, this provides an instance of learning the hard way.

First, a caution. The relevant data matrix appears in Table 15.3 (p. 376) and is reported as based on $N=250$ when in fact these statistics were based on $N=137$. Kline says some "technical problems" (p. 375) were avoided or resolved by *artificially* increasing N from 137 to 250, but he provides no indication of the nature of the resolved problems. Artificial increases in N are disconcerting and should be discouraged, but this seems likely a mere indiscretion in comparison to another feature of Kline's Figures 15.3 and 15.5 models that is likely to be unjustifiably emulated. Kline includes, but never defends, why these models permit correlation between successive (time-adjacent) measurement error variables. It is easy—too easy, and too easily emulated—to contend that the mere proximity of one measurement to the next warrants measurement error covariances. Omitting these error covariances results in model failure (Table 15.4), but even with the error covariances included, the models remain significantly inconsistent with the data. Kline "retained" both models, thereby persisting in his troubling disregard for model test evidence (p. 380, 385), but his inclusion of dubious error covariances to transform a highly data-inconsistent model into a model displaying modest but still significant ill fit enticed me to consider this more carefully.

I used LISREL to replicate the results reported in Kline's text and website, and then altered the model in a way I considered to be more theory-defensible. I thought each subject's real performance on the air traffic control learning task at any one time would influence their subsequent performance. That is, I viewed each participant's true and improving performance at any one time as likely to persist and contribute improved performance on their next attempt at the task. This conceptualization recommends replacing Kline's five error covariances with five effects, leading from the performance on each trial to the performance on the next trial. The result was that my version of Kline's Figure 15.3 model fit (χ^2=6.422, *df*=7, *p*=.492), while Kline's model did not (χ^2=16.991, *df*=7, *p*=.017; p. 381); and my version of Kline's Figure 15.5 model fit (χ^2=7.886, *df*=11, *p*=.724), while Kline's model did not (χ^2=27.333, *df*=11, *p*=.004; p. 381). These χ^2 values use Kline's artificial *N* of 250, and my models' fits were further "improved" by using the proper *N*=137. I encountered no unusual estimation difficulties using the real *N*, and I encountered no sign of technical problems that might have warranted using Kline's artificial *N*=250.

My models closely reproduced the pattern of means in the data, and confirmed the relevance of the "Ability" variable in Figure 15.5, but provided somewhat different explanations for how the observed means are accounted for via Kline's Initial and Shape latents. Kline knows that in the context of latent growth curves, earlier observations of values of a variable can sometimes influence later values (see his p. 391 figure), though my models recommend retracting his claim that effects leading to successive observations and questions regarding latent growth curves "cannot be answered in the same model" (p. 392)—because that is precisely what my models did. I should also report that including measurement error variance in the Ability variable further improves my model, and would presumably improve Kline's Figure 15.5 model if he included measurement error variance on this single indicator.

The gist of this story is that Kline was caught in the act of inserting undefended error covariances merely to reduce χ^2 ill fit, when in fact a substantively reasonable and cleanly fitting model was easily attainable. The warning is clear: Do not insert coefficients merely to improve fit. The alternative is equally clear: Pay attention to the data-providers' causal world. Readers are also encouraged to consider why Kline failed to detect his problematic model specification despite employing the two-step approach that was supposed to make "it easier to detect potential specification error" (p. 376–77).

Chapter 16, Multiple-samples analysis and measurement invariance

Multi-group analyses are most commonly used to assess measurement invariance, which Kline illustrates using both continuous and ordinal indicators. Additional uses for multi-group analyses are granted a paragraph of discussion (p. 395) but readers are not informed about how such models can be used to: identify otherwise underidentified models, control for variables that are unmeasured in some groups, or to integrate complementary but non-overlapping model segments from diverse data sets (Hayduk 1996: Chapter 5).

As in the preceding portions of Kline's text, the writing is effective and informative but also laced with model-testing laxity and the occasional oddity. Kline's section on "testing strategy and related issues" reports that "Failure to retain the invariance hypothesis at a particular step means that even more restrictive models are not considered" (p. 399–400), but Kline fails to explicitly report that this testing requirement begins at the very first step (namely, with the baseline configural model) because χ^2-difference testing (such as for equality of coefficients between-groups) is only statistically justified if the sampling distribution for the less-constrained model has a χ^2 distribution. Hayduk (2016) underscores this point and illustrates a way to strengthen baseline configural model assessments. Fortunately, Kline's exemplified baseline configural model fits the data. An oddity is that Kline rejects this fitting model and adds a coefficient (on the basis of an ill fit covariance and without substantive justification) to further improve the model fit (p. 406, Table 16.2 and website). Kline's next example also inserts a coefficient merely to improve fit but reverts to "retaining" failing models (p. 416–17, Table 16.6).

Chapter 17, Interaction and multi-level modelling

This chapter begins with an effective presentation of interaction with observed variables, but continues Kline's waffling between regression models and causal models (p. 424–29). Causal requirements emerge by page 432, but Kline's unwillingness to routinely encourage and require causal structuring becomes uncomfortably obvious when later on that page he applies the term "cause" when time-sequencing is available and "effect" for cross-sectional designs. Unfortunately time sequencing is not an appropriate criterion because reciprocal effects can be estimated with cross-sectional data and both the reciprocally connected variables can't possibly be "first." Urging use of the terms *cause without effect*, and *effect without cause*, along with a dubious differentiating criterion, could be called self-inflicted befuddlement.

My following comments aim to improve the interpretation of mediated interactions (as discussed on p. 435–37) by expanding Kline's Figure 17.5c example to illustrate some general principles and important overlooked complexities. Figure 4 corresponds to Kline's Figure 17.5c, with coefficients from his Equation 17.7. Intercepts are represented with dashed lines and disturbances/errors are designated as e's. Kline expressed the model as "two unstandardized *regression* equations" (Equations 17.7; p. 435, emphasis added). I express these as putatively causal equations by replacing the regression-predicted $\hat{Y}$ values with Y and the relevant causal-disturbance/error e_Y, and by similarly converting the mediator-moderator "regression" equation into a causal equation by replacing $\hat{M}$ with M and disturbance/error e_M.

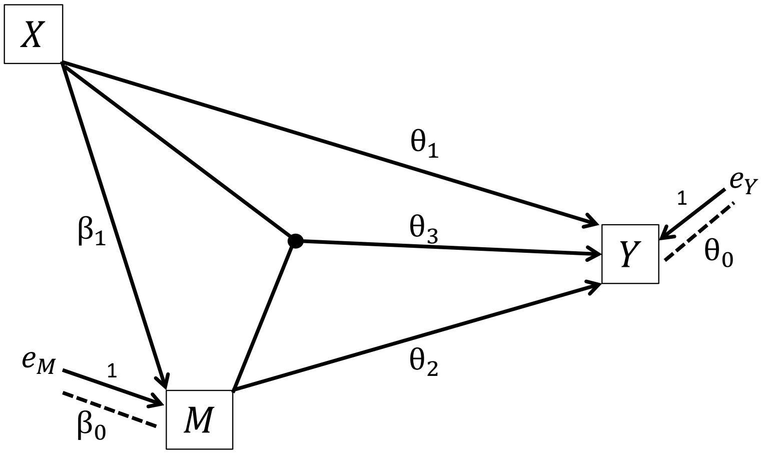

Figure 4. ***Kline's Figure 17.5c, expanded and re-expressed with coefficients.***

$$M = \beta_0 + \beta_1 X + e_M \tag{16}$$

$$Y = \theta_0 + \theta_1 X + \theta_2 M + \theta_3 XM + e_Y \tag{17}$$

The first model equation reports X as a cause of mediator/moderator M. Since M causes Y in the second equation, M is postulated as mediating part of X's causal impact on Y. M also functions as a moderator (interacting variable), represented by M's product with X in the second equation. β_0 and θ_0 are intercepts capturing the net impact of variables currently excluded from the equations. Before returning to Kline, we rearrange these equations by inserting the first equation into the second (which corresponds to replacing M with M's causal foundations):

$$Y = \theta_0 + \theta_1 X + \theta_2(\beta_0 + \beta_1 X + e_M) + \theta_3 X(\beta_0 + \beta_1 X + e_M) + e_Y \tag{18}$$

and multiplying out while inserting { } to keep track of the origin of the various terms:

$$Y = \theta_0 + \boldsymbol{\theta_1 X} + \{\theta_2 \beta_0 + \boldsymbol{\theta_2 \beta_1 X} + \theta_2 e_M\} + \{\boldsymbol{\theta_3 X \beta_0} + \boldsymbol{\theta_3 X \beta_1 X} + \boldsymbol{\theta_3 X e_M}\} + e_Y \tag{19}$$

Each term in this equation provides a causal constituent of *Y*, and the bolded terms document where and how the *X* variable participates in producing *Y*.

Supplement Section 6 considers each right-hand term to determine what the model claims would be the consequences of an intervention changing a treatment *X* from 0 (no treatment) to 1 (treatment). Each term in the expanded equation is examined to see whether and how it responds to the postulated intervention. The relevant interpretation consists of whatever wordings accurately describe the right-hand terms changing as a consequence of *X* changing from 0 to 1, and the composition of each term details the causal features providing that component of the effect transmitted to *Y*. The required assumptions consist of wordings reporting features required to render some right-hand terms constant and hence unable to produce change in *Y* as *X* changes from 0 to 1. A model's implications for a postulated intervention have been thoroughly considered if the researcher examines the coefficients and variables comprising all the terms in *Y*'s expanded equation, and reports the assumptions/presumptions required to attain and respect the intervention of interest.

Kline does not report equations corresponding to Equations 18 or 19 and instead moves directly from the model equations (like Equations 16, 17) to a set of equations reporting how to calculate the effects of intervening to change *X* from 0 to 1. Supplement Section 6 follows the procedure outlined above and locates some unacknowledged assumptions and requirements of Kline's effect calculations. For example, Kline's equations are inappropriate if the treatment happened to be coded 1 = no treatment and 2 = treatment instead of 0 and 1, and his equations do not apply if there are two treatment levels so that 0 corresponds to no treatment, 1 to weak treatment, and 2 to strong treatment. And Kline's formulas apply only to this specific model and one specific intervention. The Supplement Section 6 procedure of beginning with the equation for the dependent variable of interest, and replacing each moderator variable in the equation with that moderator's causal sources, is applicable to a wide variety of models and can examine a diverse range of potential interventions.

As it stands, researchers with more complex conditional models are cornered into trying to squash descriptions of their model into Kline's wordings rather than having been equipped to develop interpretations appropriate for their particular model. Kline's text leaves readers ill-prepared for considering consequences of interventions unavoidably making two simultaneous changes, or the consequences of reducing or increasing some effect (without intervening to change any variable), or assessing which specific disturbance/error variable's values might disrupt or assist the causal impact of interest. Assessing such interventions becomes feasible using the procedure illustrated in the Supplement.

Just as the interaction/moderator segment of Chapter 17 could be improved by considering model equations (as above), the chapter's discussion on multi-level modeling could be similarly improved. Kline's figures, Mplus syntax, and output (on the publisher's website) are appropriate but the basic principles and model details remain obscure without the model equations. Equations would clarify why s1 and s2 appear in one portion of Figure 17.7c as effects and in another portion as variables. And model equations would clarify why "Game" is boxed in one portion of the figure and circled in another. Currently the reader is left puzzling how slopes and intercepts can be variables, and can be justifiably perplexed by noticing slopes are designated s1 and s2 while the intercepts seem to be AWOL. How is a reader to understand why the same indicator variables appear in two parts of Figure 17.8b, and determine whether the disturbances/errors on the duplicate indicators are the same? Clearly there are too many potholes for this review to fill, though the road to improvement is paved with equations.

Chapter 18

Kline's concluding chapter accumulates and effectively structures the recommendations and advice provided in earlier chapters, and hence it reflects *Practices* but not "*Best Practices" in Structural Equation Modeling*. The chapter begins by tabling several references offering suggestions on conducting and reporting SEM studies. The table's footnote reports the third of ten SEM commandments as: "simpler models are better" (p. 453). A simple model of a moderately complex world is likely to be misspecified, so surely "Best Practice" would recommend an appropriately complex model, not merely a simpler model!

"Best" requires acknowledgeing that some practices as better than others, and support for the stronger practice. Unfortunately, if we consider model testing, Kline continues to promote weak practice. Model testing is not even granted its own section in Chapter 18, and it is mentioned as only one of 16 points under the heading *Estimation*. Even there the wording "Never retain a model based solely on global fit testing" (p. 461) is slanted to suggest retention of test-failing models, as opposed to respecting evidence and pursuing the sources of detected model-data inconsistencies.

Kline's Chapter 18 sections on model *Specification* (p. 454) and *Respecification* (p. 463) could similarly be strengthened by encouraging consistent pursuit of models mirroring the world's causal structure. Structuring models to reflect specific theories is laudable but limited. A researcher committed to a theory-based model that demonstrates data-inconsistency will flounder until they reground themselves in the quest for understanding the world by seeking a new or modified theory. Consider the risk created by routinely including residual/error correlations in models (recall the problematic models from Chapter 15, p. 378, 385), and the risk arising from attempting to fix failing models by adding coefficients suggested by modification indices or specific residuals (recall the problematic negative estimate p. 347). The risk is *not* merely of "capitalizing on chance" (p. 455), or that this constitutes an exercise in "chasing sampling error" (p. 463), or that this incurs a "cost of too many parameters" (p. 463). This risks obscuring (by incorrectly modeling) real stable evidence that is inconsistent with the substantive structure of the current model. Inserting coefficients merely as a matter of "routine" or because the coefficients have large modification indices risks inserting worldly-inconsistent coefficients that absorb and obscure whatever real data covariance inconsistencies managed to speak against the original theory/model. The fundamental risk is that real (not merely sampling error) covariances are modeled in the wrong way. Replication will not detect or correct models based on improper "causal" accounts of real covariances. The real covariances remain stable and so the researcher is condemned to proceeding with a now-fitting-and-replicating but nonetheless wrong model, and hence is robbed of the opportunity to get the model right.

Lack of commitment to seeking the world's structure is also evident when Kline says "do not specify feedback loops as a way to mask uncertainty about directionality" (p. 455). The options Kline leaves open to this researcher are to choose one causal direction or the other, or to drop both effects. That is, Kline implies the researcher should include a non-theory based effect-directionality (or gap) into their model, rather than encouraging the researcher to introduce exogenous causes that would make the reciprocal effects estimable and thereby permit the worldly-data to potentially support the existence of both, either, or neither of the theory-eluding reciprocal effects.

Similarly, consider the flaccid commitment to seeking a world-matching model in the context of measurement. "Multiple-indicator measurement is generally better than single-indicator measurement" (p. 454). If the researcher begins by providing each latent the best available indicator, each additional indicator is prone to being weaker and more problematic, and hence more indicators does not necessarily constitute better modeling. "Best practice" would begin with the best indicators and supplement these with only strong additional indicators. Two or three indicators per latent substantially increase model testing power (Hayduk and Littvay 2012) and are likely to be sufficient to detect specification problems—presuming the researcher respects model-test evidence of problems. Multiple indicators unavoidably include weaker indicators, and expand models in ways which tend to squeeze out latents clarifying mechanisms of action or contributing informative controls—which results in generally worse, not better, models.

And consider the claim that a way to improve on a single indicator is "to specify an instrument for the single indicator" (p. 454). A reader would be justifiably mystified by how a problem with a single-indicator is to be overcome by introducing another single-indicator (namely the instrument) into the model. By downplaying the relevance of the world's causal structuring, Kline is cornered into expressing this as if the improvement somehow comes from statistics (the statistics of instrumental variables) rather than from employing single indicators in ways that benefit from, and capitalize on, the world's causal structure.

Turning to *Identification* (p. 457), it is reasonable to check that the number of data covariances exceeds the number of estimated coefficients, and the identification of simple models should indeed be checked. But the unavailability of general procedures for checking full or moderately complex models should have prompted suggestions for: locating likely problematic model segments, checking maximum likelihood iterations (if maximum likelihood estimation is used), checking for unexpected estimate signs or magnitudes, and checking for inflated standard errors. Solutions to underidentification should also have been included—namely adding data constraints (e.g. additional identification-helpful variables) or adding model constraints (e.g. fixing/specifying or constraining model coefficients). Archival data may offer fewer opportunities to improve identification by adding relevant variables, but it is simply a mistake to claim "that the model is not identified" (p. 459) merely because it is based on archival data. Fixing/specifying coefficients to attain identification is particularly relevant with archival data, especially if the researcher investigates the sensitivity of the model to a realistic range of fixed coefficient values—including non-zero values for unresolved latent-level loop or reciprocal effects.

Model *Respecification* (p. 463) provides another instance where it would be helpful to differentiate between fitting models and proper models. Consider a researcher in a discipline confronting worldly causal structures that are not yet understood. If the researcher's model fails, the basic options are: add coefficients that reduce the model's ill fit, report model failure, or junk the model. Junking the model seems a waste, and reporting failure of a model is likely to be personally uncomfortable, so the common response is to add coefficients until the model's fit can be passed off as good enough. Unfortunately, as long as model respecification is touted as being a matter of each particular model's local or global fit, the respecification is likely to fall short of addressing the deeper disciplinary concerns. The concern is not that the "good fit is achieved at the cost of too many parameters" (p. 463). The concern is that even one additional coefficient may be sufficient to obscure the evidence recommending that the discipline reconsider the thought modes underlying the whole model.

Your first edition, or Kline's fifth edition

I have been unable to convince myself of the source(s) of Kline's reticence to notice and address the multiple and diverse concerns discussed above. To see how Kline thinks about these matters, and to glean a hint of his intentions, I requested that the editor of *Canadian Studies in Population* (Frank Trovato) invite Rex Kline to respond to this review essay. It would be nice if Kline plans a fifth edition, but if this is not planned, I hope that readers will consider preparing their own first edition, or possibly one co-authored with Kline. Irrespective of who writes the next edition, I would suggest a title like *Principles Nurturing Best Practice in Structural Equation Modelling*, where the text begins by focusing on structural equation models as striving for correct causal representations (a commitment which differentiates SEM from regression) and complementing this with routine attention to detecting and resolving model misspecification (not merely seeking fitting models). Whether or not you are the new-author, you can do SEM a service by inserting a reference to this review in whatever copies of Kline's fourth edition you encounter.

Acknowledgement

I thank Mike Gillespie for his helpful comments.

References

Anderson, J.C., and D.W. Gerbing. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin* 103:411–23.

Appelbaum, M., H. Cooper, R.B. Kline, E. Mayo-Wilson, A.M. Nezu, S.M. Rao, and C. Clinic. 2018. Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist* 73(1):3–25.

Browne, M.W., and R. Cudeck. 1992. Alternative ways of assessing model fit. *Sociological Methods and Research* 21(2):230–58.

Browne, M.W., and S.H.C. DuToit. 1991. Models for learning data. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions,* edited by L.M. Collins and J.L. Horn. Washington: American Psychological Association, p. 47–68.

Browne, M.W., R.C. MacCalum, C.T. Kim, B.C. Andersen, and R. Glaser. 2002. When fit indices and residuals are incompatible. *Psychological Methods* 7:403–21.

Box, G.E.P. 1976. Science and statistics. *Journal of the American Statistical Association* 71:791–99.

Entwisle, D.R.E., L.A. Hayduk, and T.W. Reilly. 1982. *Early Schooling: Cognitive and Affective Outcomes.* Baltimore: Johns Hopkins University Press.

Hancock, G.R., W-L. Kuo, and F.R. Lawrence. 2001. An illustration of second-order latent growth models. *Structural Equation Modeling* 8(3): 470–89.

Hayduk, L.A. 1985. Personal space: The conceptual and measurement implications of structural equation models. *Canadian Journal of Behavioural Science* 17(2):140–49.

———. 1987. *Structural Equation Modeling with LISREL: Essentials and Advances.* Baltimore: Johns Hopkins University Press.

———. 1990. Should model modifications be oriented toward improving data fit or encouraging creative and analytical thinking? *Multivariate Behavioral Research* 25(2):193–96.

———. 1994. Personal space: Understanding the simplex model. *Journal of Nonverbal Behavior* 18(3):245–60.

———. 1996. *LISREL Issues, Debates, and Strategies.* Baltimore: Johns Hopkins University Press.

———. 2014a. Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement.* 74(6):905–26.

———. 2014b. Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology* 14:124. DOI: 10.1186/1471-2288-14-124.

———. 2016. Improving measurement-invariance assessments: Correcting entrenched testing deficiencies. *BMC Medical Research Methodology* 16:130. DOI 10.1186/s12874-016-0230-3.

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

Hayduk, L.A., and D.N. Glaser. 2000b. Doing the four-step, right-2-3, wrong-2-3: A brief reply to Mulaik and Millsap; Bollen; Bentler; and Herting and Costner. *Structural Equation Modeling* 7(1):111–23.

Hayduk, L.A., and L. Littvay. 2012. Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology* 12:159. DOI 10.1186/1471-2288-12-159.

Hayduk, L.A., R.F. Stratkotter, and M.W. Rovers. 1997. Sexual orientation and the willingness of Catholic seminary students to conform to church teachings. *Journal for the Scientific Study of Religion* 36(3):455–67.

Hayduk, L.A., H. Pazderka-Robinson, G.G. Cummings, M.D. Levers, and M.A. Beres. 2005. Structural equation model testing and the quality of natural killer cell activity measurements. *BMC Medical Research Methodology* 5(1):1–9 plus corrigenda of Table 1, value .922 to .992, and .944 to .994. DOI10.1186/1471-2288-5-1.

Houghton, J.D., and D.L. Jinkerson. 2007. Constructive thought strategies and job satisfaction: A preliminary examination. *Journal of Business Psychology* 22:45–53.

Kanfer, R., and P.L. Ackerman. 1989. Motivation and cognitive abilities: An integrative/aptitude–treatment interaction approach to skill acquisition. *Journal of Applied Psychology* 74(4):657–90.

Kenny, D.A., and C.M. Judd. 1984. Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin* 96(1):201–10.

Kline, R.B. 2011. *Principles and Practice of Structural Equation Modeling*. 3rd edn. New York: Guilford Press.

———. 2013. *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*. 2nd edn. Washington: American Psychological Association.

Mulaik, S.A. 2009. *Linear Causal Modeling with Structural Equations*. Boca Raton, FL: Chapman and Hall/CRC.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge (UK): Cambridge University Press.

Rigdon, E. 1995. A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research* 30(3):359–83.

Valeri, L., and T.J. VanderWeele. 2013. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macro. *Psychological Methods* 18(2):137–50.