DOI 10.25336/csp29418
Canadian Studies in Population 45
no. 3–4 (2018), p. 188–95

Response to Leslie Hayduk's review of *Principles and Practice of Structural Equation Modeling*,[1] 4th edition

Rex B. Kline[2]

I want to thank Dr. Leslie Hayduk—Les, from this point on—for his detailed commentary on the fourth edition of my *Principles and Practice of Structural Equation Modeling*. I called Les to thank him in person just before I wrote this response. Les is a valued colleague who is also known for his depth of knowledge, strong appreciation for evidence, and forcefulness in expressing his viewpoints about the practice of structural equation modeling (SEM). I still have my copy of Les' book on SEM with LISREL (Hayduk 1987), which was one of the first textbooks, if not the very first, in the area. I cut my SEM teeth, so to speak, on Les' book at the beginning of my career. I also want to thank the *CSP* editor, Dr. Frank Trovato, for the chance to respond to Les' review. I've never had the opportunity to comment on a review before it is published, and this experience has been both stimulating and rewarding.

As I tell our honours thesis students, one of the most precious gifts an author can receive is detailed criticism written by someone who knows a lot about the area *and* invests the effort to explain his or her feedback in detail. I also remind them that (1) writers need a thick skin; and (2) there's little point to becoming defensive in response to criticism, especially when that commentary is intended as constructive. This is also my view of Les' critique, which is impressive in both its breadth and length. Indeed, the latter makes it more like a chapter (or two) than an article. Because my response is not a point-by-point rebuttal, it is not nearly as long as Les' review. Instead, my comments are organized around the six themes:

1. Goals and purposes of the book.
2. Special needs of the target audience.
3. Book organization and language.
4. Comments on model fit assessment.
5. Selective responses to some other of Les' points.
6. Plans for the fifth edition.

Goals and target audience

From the very first edition in 1998, *Principles and Practice* has been written for SEM newcomers who do not have strong quantitative backgrounds. Such readers deal much more with applied research problems than theoretical topics in statistics. They also bring to learning about SEM the handicaps listed next and discussed afterward:

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 Psychology, Concordia University, L-PY 151-6 Psychology Building, 7141 Sherbrooke W., Montreal, QC H4B 1R6; email: rex.kline@concordia.ca.

1. Many, if not most, are unfamiliar with linear algebra, and thus, presentations based on matrix symbolism are not helpful.
2. They have little, if any, background in psychometrics. This means that many beginners lack the formal skills to (a) select the best measure among alternatives and (b) evaluate score precision, or reliability.
3. Newcomers often have a strong significance-testing mentality; that is, they assume that outcomes of significance testing, or *p*-values, are a scientific gold standard, or at least a decision criterion. They also assume that significance testing has the same role in SEM as in more conventional statistical techniques, such as multiple regression.
4. They believe that the goal of SEM is to *find a model that fits the data*. But this outcome has little if any meaning. This is because *any* model, even one that is grossly wrong, can be made to fit the data simply by adding effects to the model, or making it more complex. As the model becomes more complex, its fit should then improve. Instead, the *real* goal of SEM is to *test a theory* (Hayduk et al. 2007); thus, retaining no model is a perfectly acceptable outcome (e.g., p. 118).

Organization and language

Beginners require language or pedagogy that does not always satisfy those who are already expert, and it is relative easy for experts to overlook the gulf between them and novices. David Kenny mentions the same issue in the preface to his classic book, *Correlation and Causality*. After describing efforts to avoid confusing the beginner by emphasizing standardized solutions and occasionally blurring the distinction between parameters and statistics, Kenny notes, "If these practices disturb you, I apologize. But I felt that if I had to sacrifice elegance for the experts in order to obtain clarity for the beginner, I would choose clarity" (1979: ix).

So too have I made similar choices. As Les noted, the 4th edition features greater separation in coverage of *classical path models with single-indicator measurement* versus *latent variable models with multiple-indicator measurement*. Specifically, model specification and identification are introduced for path models (Chapters 6–8) before dealing with these topics for latent variable models (Chapter 9–10). This approach dulls the elegance of SEM as a single framework that accommodates both single- and multiple-indicator measurement, which was the approach taken in the 3rd edition. But dealing with both kinds of models at once can be intimidating for beginners, many of whom are more familiar with statistical techniques for observed variables. Thus, based on reader feedback, including suggestions from instructors of multivariate or SEM courses, I returned to the "old way"—that is, from the 2nd edition—of organizing the discussions for these topics. Less elegant, but more practical oomph: a worthwhile trade off, I think.

The organization just described also better supported the inclusion in the 4th edition of a new topic, Judea Pearl's structural causal model (SCM), or a graph-theoretic approach to SEM, also called third generation SEM (Grace et al. 2012). Better known in computer science and epidemiology than in psychology and related disciplines, the SCM brings new capabilities to the SEM family. One is the formal analysis of nonparametric causal diagrams, especially of a directed acyclic graph (DAG) with no casual loops. These graphical methods can be used to determine whether a particular casual effect in the DAG is identified and, if so, whether multiple estimators of that effect are available. There are freely available computer tools and calculating Web pages for this purpose, too.

Les is right in that some concepts from the SCM are not covered in the 4th edition, including Pearl's *do*-calculus, which has implications for the identification of causal effects in nonparametric models and mediation analysis, among other topics (Pearl 2014). But my goals in this area were both ambitious and limited at the same time. The 4th edition of *Principles and Practice* was one of the first introductory-level books in the "traditional" (i.e., social science) SEM literature to cover the SCM. These concepts would not only be novel for many readers, they can also illuminate key insights on casual modeling, such as identification, which is one of the most challenging problems in the conduct of SEM (Kenny and Milan 2012). That's the ambitious part. But only so much can be done in this area without overwhelming readers, and I decided that covering do-calculus was just too much. That's the limited part. Sometimes it is enough to make readers aware of previously unknown possibilities that can be explored later.

Les is also unhappy with my use of language about analyses where means are analyzed along with covariances. These analyses feature a constant that is regressed on exogenous or endogenous variables in the model in order to estimate, respectively, means or intercepts. For observed variables, these analyses with the constant are carried out exactly as in computer procedures for multiple regression when intercepts are calculated (e.g., p. 369–74). The diagram symbolism I use in the book comes from the McArdle–McDonald reticular action model (RAM; McArdle and McDonald 1984). In RAM symbolism, every parameter is explicitly represented with its own special symbol in model diagrams. This feature helps beginners to understand all is going on in a particular analysis.

In analyses where means are analyzed, the constant in RAM symbolism is represented in the model diagram as having direct or indirect effects on observed or latent variables. I tell readers that these effects do *not* have the usual interpretation because the constant is not a variable, and thus is not causal (p. 371). Nevertheless, applying the tracing rule to a model diagram with the constant in RAM symbolism allows the reader to see exactly how predicted means or intercepts are derived in the analysis. For example, the sum of all direct or indirect paths from the constant to an endogenous variable is a predicted mean. In the output for some SEM computer tools, such as EQS, model-implied intercepts or means are represented in effect decompositions as, respectively, direct versus total effects of the constant. I appreciate that Les objects to use of term "effects" when describing the constant, but I believe this inelegant use of language has explanatory power. Also, I find RAM symbolism for the constant to be easier to understand for beginners than Les' suggested alternative (see his Figure 3).

The measurement crisis in psychology refers to a substantial decline in the quality of instruction about psychometrics over the last 30 years or so, during which courses on measurement disappeared from many undergraduate and graduate programs (Lambert 1991). The same term also describes the widespread failure of too many researchers to estimate and report the reliabilities of scores analyzed (Vacha-Haase and Thompson 2011). A consequence is that many SEM newcomers have little knowledge of psychometrics, yet analyses of latent variable models with multiple-indicator measurement requires strong skills in this area. This is why part of the chapter on data preparation in the 4th edition is devoted to the basics of psychometrics and measure selection. Many readers sorely need guidance in these crucial areas.

A related topic is the method outlined in *Principles and Practice* that serves as an alternative to representing a single indicator as one would in a classical path model (p. 214–17). Briefly, in this method (1) the observed variable is specified as the sole indicator of a latent variable; (2) the error variance is fixed to equal the quantity $1 - r_{XX}$ times the sample variance of the indicator, where r_{XX} is a reliability coefficient; and (3) the unstandardized pattern coefficient (factor loading) for the single indicator is fixed to 1.0, which scales the corresponding factor.

The single-indicator method just described is simple to implement and does not change overall model fit. It also provides a way to deal with a major limitation in standard multiple regression analysis, the assumption that scores for all predictors are perfectly precise, or r_{XX}=1.0 for each and every predictor. If multiple predictors are measured with error plus there is error on the criterion (for which r_{XX} = 1.0 is *not* assumed), the results can be biased to a degree beyond the expectations of probably most researchers who use multiple regression. For example, results of significance testing about incremental validity can be very untrustworthy in the presence of even modest amounts of measurement error (Westfall and Yarkoni 2016). Results of analyses of classical path models can also be severely biased, if measurement error is not taken into account (Cole and Preacher 2014).

Les pointed out that it is not always necessary to fix the error variance for a single indicator, and his Figure 2 depicts situations when direct effects from a latent variable help to identify the error variance. This is a good point. My larger goal is to promote awareness among readers of the possibility to explicitly represent measurement error even in models with observed variables only. If they do so using any variation of methods for single indicators that control for measurement error, then (1) at least a small part of the measurement crisis is addressed, and (2) I would be happier as a result.

Assessing model fit

The significance testing crisis refers to the ongoing debate, now occurring in many disciplines, about the proper role of significance testing, *if any*, in data analysis (Gelman 2018; Szucs and Ioannidis 2017). Significance testing is now banned in the journal *Basic and Applied Social Psychology* (Trafimow and Marks 2015), and the American Statistical Association issued a statement warning against misuses of *p*-values (Wasserstein and Lazar 2016). There is also ample evidence that most researchers do not understand *p*-values (Kline 2013). The collective effect of multiple cognitive errors about *p*-values is confirmation bias; that is, researchers believe that sample data supports their hypotheses to an extent that far exceeds reality.

Most students and researchers in the social sciences are steeped in the mythology of significance testing, and this background can interfere with new learning about SEM. For example, beginners in SEM generally endorse the false beliefs listed next:

1. Statistical significance for individual path coefficients, such as $p < .05$, is evidence that the model is correct.
2. Paths with coefficients that are not significant must be dropped from the model.
3. Significant modification indexes for parameters not yet specified as free signal missing truths; that is, such effects should be added to the model.

Altogether the false beliefs just listed promote the retention of models that massively capitalize on sampling error, and thus are unlikely to replicate. But things get even weirder in significance testing land when it comes to model evaluation, as explained next.

A widespread but poor practice in SEM occurs when researchers otherwise preoccupied with *p*-values for tests of individual parameter estimates *ignore* the outcome of model χ^2 test for the whole model. This logical contradiction is motivated in part by the false belief that the model χ^2 statistic is affected by sample size in all situations. As Les noted—and I tell readers as much (e.g., p. 271)—the model χ^2 is affected by sample size only when the model is incorrect, or belongs to an equivalence class of models that do not imply the sample covariance matrix. Ignoring a significant

model χ^2 is bad practice especially if the power of the test is low, which is typically true in most published SEM studies (Wolf et al. 2013).

If a model fails the χ^2 test under conditions of low power, then the degree of misspecification could be severe. Without investigating further—that is, diagnosing model–data correspondence at the level of the residuals—the researcher might falsely conclude that the model fits the data. But the χ^2 test can be fooled, too, such as when power is low. For example, certain residuals could be relatively high, which signals poor fit at the level of pairs of observed variables, in the presence of a model χ^2 that is not significant. Also, it is easy to get a model χ^2 that is not significant just by freeing additional model parameters, or making a model more complicated. If there is little justification in theory for the added complexity, then the respecified model may not replicate due to extreme capitalization on chance.

In Les' world, the model χ^2 is the only acceptable global fit statistic. Specifically, he would reject the use of any other global fit statistic, especially approximate fit indexes (Hayduk et al. 2007), such as the root mean square error of approximation (RMSEA), among others. Approximate fit indexes are not significance tests; instead, they are intended as continuous measures of model–data correspondence and in this way are analogous to effect size statistics. Unfortunately, a mythology about approximate fit indexes has grown along with the size of this family of global fit statistics. The primary myth is that there are threshold values of certain approximate fit indexes that can reliably differentiate between models with "good" fit versus those with "bad" fit. The same myth is behind the poor practice of retaining a model that has failed the χ^2 test and without inspecting the residuals based on the observation that values of particular approximate fit indexes exceed their respective thresholds. I condemned this poor practice in pretty clear terms (Chapter 12).

In Rex's world, the reporting of values of certain approximate fit indexes is acceptable if (1) there is no reference to magical cutting points for such statistics, and (2) the researcher also reports on the residuals, which are the details of fit. I think there are certain approximate fit indexes, including the RMSEA with its 90% confidence interval, that are so widely reported that reviewers would be suspicious if this information were omitted. Thus, I advise readers to report values for a core set of approximate fit indexes but not to base the decision about whether to retain the model solely on global fit statistics of any kind, including the χ^2 test. I think Les is appalled by this stance, even though I agree with his criticisms of approximate fit indexes.

Because global fit statistics of any kind provide rather crude information about average or overall model fit, I emphasize analysis of the residuals and also reporting on such analyses in written reports (e.g., p. 254, 311, 330, 346–47, 380, 385, 408, 417). Unfortunately, severe misspecification is not always obvious in the residuals, but ignoring the residuals in SEM in analogous to ignoring the residuals in regression analysis. There are some important differences between the two: residuals in SEM are typically at the level of pairs of observed variables, but regression residuals are calculated for individual cases. But it would nevertheless be just as foolhardy to ignore regression residuals as it would be to pay no attention to the residuals in SEM. I find this sorry practice in SEM to be appalling. But I think that Les and I would agree that typical practice about fit assessment in the empirical SEM literature is poor, although we emphasize different things. Les may see those differences as irreconcilable, while I am less pessimistic.

Other comments

Next I will address a few other items in Les' review. Many of the examples in *Principles and Practice* are secondary analyses of data collected by other researchers. A limitation is that it is not always possible to specify a priori hypotheses about which particular parameters should be fixed versus freed in model respecification. This is why in some cases I did not decide about which

among alternative respecifications would be the best when no model is retained (e.g., p. 285–86, 312). In other cases when, for pedagogical reasons, I add particular effects to the model, I try to explain the bases for these decisions. Perhaps even more warning about the context is needed; that is, a pedagogical example is not wholly representative of model testing by researchers who are experts in that area.

In some cases I made decisions about model specification that differed from those of the original authors (e.g., p. 341–48, 408). In each case I offered explanation, but I do not claim that my decisions were actually better. Also, it can and does happen in the complex multivariate analysis that different researchers will make somewhat different decisions. It is best to be open about the role for discretion in statistical modeling and also to alert readers that such decisions require explanation. One hopes that the findings would be sufficiently robust to hold up over variations on how the analysis is conducted, but this does not always happen. Indeed, this is the point of a sensitivity analysis: the same data are analyzed under somewhat different assumptions. If the results depend on a particular set of assumptions that are not clearly preferred, given the research problem, then little confidence may be warranted in the stability of those results.

Les pointed out an example where my interpretation of an error correlation was wrong, and I thank him for pointing out this error. The correlation is for a pair of indicators in a structural regression model described by Houghton and Jinkerson (2007). The error correlation is −.243 (p. 347), and I mistakenly interpreted this result as indicating the effects of shared content over two different indicators of subjective well-being. As noted by Les, a positive error correlation would be consistent with my original explanation, not a negative correlation. I believe I mistakenly thought that the wording of the two subjective well-being scales was reversed, which would predict a negative residual correlation due to shared content, but that is not the case.

I am not surprised that Les was able at least one find an alternative model to the latent growth model described (p. 375–87). Les' alternative is a kind of autoregressive model where indicators measured at earlier times have direct effects on indicators measured later. Models with autoregressive structures can be viable alternatives to traditional latent growth models, although Little noted some exceptions (2013: 271–73). The larger point is that alternative models can include not just variations on a model within the same class of models, but also variations over types of measures, such as latent growth versus autoregressive in this example. Such alternative models may be near-equivalent models with similar, but not identical fit to the same data.

Fifth edition

Finally, will *Principles and Practice* see a 5th edition? Yes. After finishing each edition, I starting planning for the next one. At this point I can share a few ideas. The 4th edition will be the last of its kind in that it has reached the limits of its growth. This is because anything longer would be a tome, and thus less effective in getting beginners off to a good start. I think that chapters in the 4th edition about background concepts, such as the basics of regression analysis, significance testing, data screening, and psychometrics, would be updated but made available as supplemental materials to the 5th edition. Most beginners still need to review these topics, but readers with stronger quantitative background or experience in test construction would benefit less from this material.

I see Pearl's graph-theoretic approach as playing an even larger role as a framework for understanding both nonparametric causal models and more traditional (at least in the social sciences) parametric causal models. The increasing availability of computer tools that determine whether particular direct or total effects are identified and, if so, exactly how to estimate the corresponding effect through covariate selection or instrumental variables can help researchers deal with

the problem of identification. But such computer tools are less helpful in dealing with models where latent variables are specified as measured by multiple indicators, such as confirmatory factor analysis measurement models. So there is still room for old-school knowledge—that is, second generation SEM (i.e., now)—especially when testing hypotheses from the perspective of classical measurement theory.

That's enough speculation for now. Returning to the present, I want to say that (1) no book is perfect, including the 4th edition of *Principles and Practice*; and (2) different authors will have different ideas about organization, wording, and examples. That's how authorship plays out. I want to again thank Les for his extensive comments, which give me many ideas for the 5th edition. Les and I spoke about an SEM-related book he is working on now. I won't offer up any spoilers, but I think a book of this type is needed; that is, I'll be one of first customers, if the project comes to fruition. Finally, I want to encourage younger scholars and researchers to think about replacing us dinosaurs—Les, me, and others of our age cohort—who can carry things forward only for so much longer. Those who will develop and refine fourth-generation SEM are still at early points in their careers. In the meantime, I look forward to continued work in the area and ongoing interactions with students and colleagues. Blessings to all.

References

Cole, D.A., and K.J. Preacher. 2014. Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods* 19:300–15. https://doi.org/10.1037/a0033805.

Gelman, A. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* 44:16–23. https://doi.org/10.1177/0146167217729162.

Grace, J.B., D.R. Schoolmaster, G.R. Guntenspergen, A.M. Little, B.R. Mitchell, K.M. Miller, and E.W. Schweiger. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3(8):1–44. https://doi.org/10.1890/ES12-00048.1.

Hayduk, L.A. 1987. *Structural Equation Modeling with LISREL: Essentials and Advances*. Baltimore, MD: Johns Hopkins University Press.

Hayduk, L., G. Cummings, K. Boadu, H. Pazderka-Robinson, and S. Boulianne. 2007. Testing! testing! one, two, three—Testing the theory in structural equation models! *Personality and Individual Differences* 42:841–50. https://doi.org/10.1016/j.paid.2006.10.001.

Houghton, J.D., and D.L. Jinkerson. 2007. Constructive thought strategies and job satisfaction: A preliminary examination. *Journal of Business Psychology* 22:45–53. https://doi.org/10.1007/s10869-007-9046-9.

Kenny, D.A. 1979. *Correlation and Causality*. New York: Wiley.

Kenny, D.A., and S. Milan. 2012. Identification: A nontechnical discussion of a technical issue, in *Handbook of Structural Equation Modeling*, edited by R.H. Hoyle. New York: Guilford, p. 145–63.

Kline, R.B. 2013. *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*. 2nd edn. Washington, DC: American Psychological Association. https://doi.org/10.1037/14136-00.

Lambert, N.M. 1991. The crisis in measurement literacy in psychology and education. *Educational Psychologist* 26:23–35. https://doi.org/10.1207/s15326985ep2601_2.

Little, T.D. 2013. *Longitudinal Structural Equation Modeling*. New York: Guilford.

McArdle, J.J., and R.P. McDonald. 1984. Some algebraic properties of the Reticular Action Model for moment structures. *British Journal of Mathematical and Statistical Psychology* 37:234–51. https://doi.org/10.1111/j.2044-8317.1984.tb00802.x

Pearl, J. 2014. The do-calculus revisited, in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, edited by N. de Freitas and K. Murphy. Corvallis, OR: AUAI Press, p. 4–11.

Szucs, D., and J.P.A. Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience* 11:1–21. https://doi.org/10.3389/fnhum.2017.00390.

Trafimow, D., and M. Marks. 2015. Editorial. *Basic and Applied Social Psychology* 37:1–2. https://doi.org/10.1080/01973533.2015.1012991.

Vacha-Haase, T., and B. Thompson. 2011. Score reliability: A retrospective look back at 12 years of reliability generalization. *Measurement and Evaluation in Counseling and Development* 44:159–68. https://doi.org/10.1177/0748175611409845.

Wasserstein, R.L., and N.A. Lazar. 2016. The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician* 70:129–33. https://doi.org/10.1080/00031305.2016.1154108.

Westfall, J., and T. Yarkoni. 2016. Statistically controlling for confounding constructs is harder than you think. *PLoS ONE* 11(3). https://doi.org/10.1371/journal.pone.0152719.

Wolf, E.J., K.M. Harrington, S.L. Clark, and M.W. Miller. 2013. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement* 73:913–34. https://doi.org/10.1177/0013164413495237.