Available online at www.sciencedirect.com
ScienceDirect
Journal of Consumer Psychology 20 (2010) 221–225
Journal of CONSUMER PSYCHOLOGY

Methods Dialogue

Some things you should know about structural equation modeling but never thought to ask

Leandre R. Fabrigar*, Ronald D. Porter, Meghan E. Norris

Queen's University, Canada

Received 10 March 2010; accepted 10 March 2010
Available online 10 April 2010

Abstract

Iacobucci (2009, 2010) covers a number of important issues in the use of structural equation modeling and in so doing provides researchers with many useful insights and sensible suggestions. This commentary focuses on three issues where our views differ somewhat from those expressed in the target articles: SEM and causal inferences, sample size, and model fit. In addressing each of these issues, our perspectives do not so much contradict the views expressed by Iacobucci as they reflect a somewhat different conceptual emphasis.

There is little doubt that structural equation modeling (SEM) has been among the most influential and widely used statistical methods to emerge in consumer psychology and related disciplines over the past 30 years. As with any popular statistical method, applications of SEM in consumer psychology and other disciplines have not always followed the best available practices as suggested by the methodological literature. One reason for this gap between the methodological and substantive literatures is that many important methodological findings, by virtue of where they are published and their highly technical nature, are not readily accessible to the typical researcher. In light of this reality, articles such as those by Iacobucci (2009, 2010) provide a valuable resource to researchers.

Our goal in the present commentary is not to reiterate the many useful insights made in the target articles or to echo the numerous sensible suggestions provided by Iacobucci. Instead, we have chosen to focus our comments on some select issues where our views differ somewhat from those expressed in the articles. In addressing these issues, it is worth noting that our views do not so much contradict what is expressed in the target articles, but rather reflect a somewhat different emphasis and conceptual perspective.

SEM, the nature of data, and causal inferences

As Iacobucci notes, SEM is most often utilized in the context of non-experimental data. Despite this fact, users of SEM often interpret their results in strong causal terms. Indeed, when SEM was a comparatively new technique in the social sciences, much of the initial enthusiasm for it seemed to be driven by the erroneous belief that it could allow researchers to miraculously transcend the inferential limitations of non-experimental data. Several decades of experience have made such extreme views less common, but it is difficult to dispute that the problem to some degree remains. This criticism notwithstanding, we think several additional observations regarding SEM and causal inferences are worth noting.

First, in the strict sense of the term, we do not object to the use of the term "causal modeling" when referring to applications of SEM. Applications of SEM usually do involve the specification of models that make causal assumptions. Hence, researchers are modeling causal relations among variables. However, modeling a causal relation is not the same as proving a causal relation. Other models making different causal assumptions might provide comparable or superior representations of the data.

Second, we would argue that SEM is not inherently more vulnerable to inferential problems regarding causality than other statistical techniques and sometimes provides advantages in evaluating the plausibility of different causal assumptions. For

* Corresponding author. Department of Psychology, Queen's University, Kingston, Ontario, Canada K7L 3N6.
E-mail address: fabrigar@queensu.ca (L.R. Fabrigar).

1057-7408/$ - see front matter © 2010 Society for Consumer Psychology. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jcps.2010.03.003

the most part, statistical techniques do not in and of themselves confer the ability to reach causal conclusions (e.g., see Wegener & Fabrigar, 2000). For example, ANOVA provides no stronger basis for causal conclusions than does multiple regression or SEM. The reason it is often sensible to reach causal conclusions with ANOVA is a function of the fact that it is almost exclusively used with experimental data. If it were applied to non-experimental data, there would be little basis for strong causal inferences. However, stating that one cannot make strong causal claims with non-experimental data does not necessarily imply that all causal assumptions for a given data set are equally plausible and that researchers should interpret their data as if they were. Statistical procedures can often be used to evaluate the relative plausibility of competing causal assumptions. We believe SEM has some advantages over many other commonly used statistical procedures in undertaking this task.

For instance, one common threat to causal inferences in non-experimental data is that the hypothesized causal variable may co-vary with other plausible causal variables and it may be one of these alternative variables that is responsible for the observed effect. Such alternative causal variables can sometimes be ruled out by statistically controlling for their effects (as is often done in multiple regression). SEM affords advantages over other methods in this task because of its ability to account for random measurement error (which can lead not only to attenuated estimates of effects, but also sometimes inflated estimates). Additionally, in some cases systematic measurement error can also be directly accounted for in a SEM model. As such, SEM can often provide a more accurate estimate of the effects of a hypothesized causal variable controlling for the effects of other potential casual variables (see Bollen, 1989).

Another threat to causal assumptions in many non-experimental data sets is that it is often plausible to reverse causal assumptions among variables in a model. SEM is extremely flexible in that it readily permits researchers to test competing models that make different assumptions regarding casual directions. These models can be compared on the basis of their fit and the conceptual plausibility of their parameter estimates. If one model is found to be clearly superior to other models, a researcher might reasonably make the case that certain causal assumptions are more plausible than others for the given data set. SEM has several advantages when comparing competing models. It is much more flexible with respect to model specification than most other statistical procedures (e.g., regression, ANOVA) and thus it is much easier to specify and directly compare competing models. For example, SEM allows the researcher to simultaneously estimate all effects in the model whereas regression approaches often require the specification of numerous regression models to estimate different parts of an overall model. Another advantage of SEM is that formal indices of model fit have been developed to evaluate how well a model accounts for the data. Evaluating model fit in many other procedures is either not possible or much less well developed. Finally, the potentially greater accuracy of parameter estimates in SEM is another advantage when comparing the performance of competing models.

One final observation regarding SEM and causality is that although SEM is often regarded as a "non-experimental data analysis method", we think such a view is unfortunate. SEM can be usefully employed in many experimental contexts. For instance, tests of mediation have become commonplace in many disciplines when attempting to gain insight into the psychological processes underlying the effects of an experimental manipulation on a dependent variable. Such analyses have typically been conducted using regression-based approaches. However, as has been noted by many methodologists (e.g., Baron & Kenny, 1986), these approaches can often lead to inflated estimates of direct effects and attenuated estimates of mediated effects. When multiple-item measures of a mediator and/or dependent variable are available, using SEM to test latent variable meditational models has the potential to provide more accurate results. Along similar lines, when researchers wish to directly compare the effects of several possible mediators, SEM provides a much more flexible and effective means of directly comparing the effects of different mediators.

Another experimental context where SEM can be quite useful is in tests of moderated mediation. Increasingly, theories in consumer psychology and related disciplines make predictions regarding meditational effects that should vary under different conditions. For example, the Elaboration Likelihood Model (ELM) of persuasion postulates persuasion variables can serve distinctly different roles under varying levels of cognitive elaboration (Petty & Cacioppo, 1986; Petty & Wegener, 1999). Such predictions often imply that a given independent variable will have a direct effect on the dependent variable at one level of a moderator and be mediated at another level of a moderator. Alternatively, different mediators may be responsible for an effect at different levels of moderators. Such predictions can be readily tested using multi-group SEM analyses (see Wegener & Fabrigar, 2000).

Finally, it should be noted that anytime a dependent variable is examined across levels of an independent variable, a researcher implicitly assumes that the fundamental psychometric properties of the dependent variable are invariant across conditions. Such assumptions are almost never tested in consumer psychology experiments. However, for some independent variables commonly studied in consumer psychology, it is plausible to postulate that these independent variables might influence the properties of measures in addition to or instead of the underlying construct (for examples in the context of attitude–behavior consistency, see Fabrigar, Wegener and MacDonald (2010)). Multi-group SEM provides a methodology for formally testing measurement assumptions by directly examining the parameters of the measurement model across levels of an experimental variable. Such analyses can allow researchers to disentangle the effects of independent variables on measurement properties from the effects on the underlying latent variables that the measures are intended to assess.

In summary, we do not think that causal language of any sort is always inappropriate when using SEM. The strength with which one can make casual inferences is best thought of as a continuum. At one extreme, all casual assumptions are equally plausible. At the other extreme, only a single causal interpretation is plausible. Analytical strategies using SEM can sometimes aid the researcher in moving toward this later end of the continuum. Additionally, it can often be used quite productively in the context

of experimental data to further aid in a researcher's understanding of the processes underlying an experimental manipulation's effects on a dependent variable.

Sample size in SEM

Without a doubt, the issue of sample size is among the most commonly asked design questions regarding the use of latent variable models (e.g., exploratory factor analysis and SEM). Unfortunately, there is no easy answer to this question. Although simple rules of thumb have been proposed in the context of factor analysis (e.g., 5 participants/measured variable, Gorsuch, 1983) and SEM, these rules of thumb have comparatively little theoretical or empirical grounding (MacCallum, Widaman, Zhang & Hong, 1999). One reason for the difficulty in determining an appropriate sample size is that it depends on many considerations. Researchers should recognize that various sample size recommendations in the literature are often based on somewhat different conceptual approaches and as such it is not surprising that they might suggest different answers. By understanding the conceptual underpinnings of different approaches to determining sample size, researchers will be able to choose the most appropriate manner of determining sample size based on their current goals.

One way of determining sample size, commonly adopted in factor analysis but also relevant to more general latent variable models, is to consider it from the perspective of accurate parameter estimation (e.g., MacCallum, Widaman, Zhang & Hong, 1999; Velicer & Fava, 1998). Here, adequate sample size is defined as the number of observations needed to obtain estimates of the model's parameters that closely match the parameter values of the model in the population. Studies approaching sample size from this perspective have indicated that the sample size needed is smaller when unique variances of measured variables are low and each latent variable is represented by at least 3 or 4 measured variables. Under such optimal conditions, sample sizes as small as 100 may be adequate. However, under moderately less optimal conditions, it may be necessary to have a sample of at least 200, and under poor conditions samples of even 400 to 800 may be insufficient.

Another approach to determining sample size is from the standpoint of statistical power. In more conventional statistical procedures, power analyses are often comparatively straightforward. However, in the context of SEM, the issue is more complicated because typically there are many hypotheses that could be tested in given model. For example, one might test a hypothesis about overall model fit with respect to a specific model fit index (e.g. see MacCallum, Browne & Sugawara, 1996). When conducting such power analyses, a researcher must determine precisely what hypothesis they wish to test using the index (e.g., a test of close fit as defined by some specified numerical value of a given fit index) and what assumptions they wish to make regarding the fit of the model in the population. Moreover, it is important to note that power analyses can be conducted for many different fit indices and the results of these analyses might suggest different sample sizes. Another approach to power in SEM is to define it in terms of a test of difference in fit between two models (MacCallum, Browne, & Cai, 2006). Finally, although not really a major focus within the methodological literature, one could define power in terms of tests of specific parameters within the model. For instance, the goal might be to test whether one parameter within a model is stronger than another or whether a difference exists in a given model parameter across different samples.

In summary, determining sample size is complicated and depends on the researcher's perspective and goals. Depending on the focus of a particular research project, one of these approaches to determining sample size may be more appropriate than another. For example, if conducting an exploratory factor analysis, considering sample size from the perspective of accurate parameter estimation may make the most sense because in this context it is comparatively uncommon to have clear hypothesis tests of overall model fit or parameter estimates. However, if a researcher is interested in comparisons of model parameters across different groups, considering sample size based on the testing of specific parameters might make more sense. Finally, if the researcher has several clearly defined competing models, sample size might be best determined on the basis of power to test differences in fit between models.

Evaluating model fit in SEM

In describing model fit indices, Iacobucci raises a number of important considerations. However, we believe that there are several additional observations regarding model fit that merit discussion. First, we believe that there is a tendency by researchers using SEM to over-emphasize the use of model fit indices at the cost of other important information. For example, we have seen numerous cases in which researchers have obtained a good model fit and then pronounced the model as a good representation of the data without ever, in any detailed way, reviewing the extent to which the model's parameter estimates provide logical support for their conclusions. Indeed, we have seen cases in which researchers present models and never report the parameter estimates at all. In our view, such practices are a serious oversight. Although it is obviously essential to know about model fit, it is equally important to evaluate parameter estimates. It is entirely possible for a model to fit well, but to provide parameter estimates that are logically impossible or do not readily follow from the nature of the constructs being examined or the model's conceptual implications (see Fabrigar & Wegener, 2009).

A second aspect of model fit that merits comment is the tendency for researchers to interpret model fit in a simple dichotomous fashion. This practice is not surprising given that there has been substantial effort in the methodological literature to establish cut-points to distinguish between good and bad fit. Of course, the ability to assert that model fit is either good or bad does have some intuitive appeal. However, dichotomous cut-off scores are somewhat arbitrary and overly simplistic (Fabrigar & Wegener, 2009). Indeed, given that descriptive indices express the degree of model fit (or lack of fit) on a continuum, reducing these scores to dichotomous categories is like taking careful measurements of the height of people and then reducing these measurements to categories of tall and short. Reducing these

measures to two categories throws away valuable information. That being said, some guidelines are clearly needed to provide a frame of reference when interpreting model fit indices. To that end, we think a more useful goal in developing guidelines would be to use multiple categories to better represent gradations of model fit (e.g., good, acceptable, marginal, and poor). One example of such an approach is Root Mean Square Error of Approximation (RMSEA) which has been conceptualized by some methodologists in terms of gradations (Browne & Cudeck, 1992).

Finally, one important issue regarding fit indices that continues to challenge researchers is what to do when they do not agree? If all model fit indices agreed and indicated that the model fit was great or that it was very bad then deciding which fit indices to use would be irrelevant. Unfortunately, fit indices do not always agree. Thus, researchers are sometimes left with the somewhat daunting task of sorting through the myriad of fit indices and deciding which ones are to be trusted. As it turns out, a number of empirical studies have provided evidence suggesting that some model fit indices are better than others at differentiating good from bad fitting models (e.g., Fan & Sivo, 2005; Hu & Bentler, 1998). These studies provide a basis for recommending a more manageable subset of indices to use a basis for model evaluation. However, some caution is warranted in using such studies as the sole basis for selection of model fit indices, because it is difficult to know the extent to which the findings of these studies are idiosyncratic to certain properties of the data, of the model, or the nature of misspecifications that were tested (e.g., see Marsh, Hau, & Wen, 2004).

Thus, when evaluating model fit indices, we recommend that researchers keep in mind the different ways in which indices define fit. Most notably, model fit indices can be categorized in terms of at least two underlying distinctions. First, descriptive fit indices can be categorized as either absolute (they index discrepancy between the model and the data in an absolute sense) or incremental (they index discrepancy between the model and the data relative to another model). Second, some indices take into account model parsimony whereas others do not. In light of these distinctions, it is not surprising that indices can sometimes provide different results. The manner in which they define fit is different. Consequently, when evaluating performance of a model, researchers should understand the conceptual nature of fit indices so as to better interpret why indices do not always agree and what lack of agreement might be telling them about their model.

For example, if an absolute fit index that does not take into account model parsimony suggests good fit, but an absolute fit index that adjusts for model parsimony suggests poor fit, this might indicate that the model does a good job accounting for the data, but that it does so at the cost of parsimony. A reverse pattern between the two indices might suggest that the model is only marginally effective in absolute terms in explaining the data, but that it does rather well when one takes into account how few parameters are used to explain the data. Ultimately it may be possible to identify specific indices that are the best representatives of each category of fit indices. However, arguing for an index from one category over an index from another category may be a bit like arguing in favor of an apple over an orange. It may be difficult to say that one is better than another, but the process of comparing them could be informative for understanding the strengths and weaknesses of the model.

Conclusions

In closing, we think it fair to concede that SEM has not always been applied as sensibly as it should have been. However, there is also little doubt that the technique has greatly enriched the manner in which we conduct research in consumer psychology and other disciplines. Moreover, as expertise in using SEM increases among researchers and methodologists continue to solve major methodological challenges in the application of SEM, we think the future for SEM in consumer psychology will be even brighter than its past.

Acknowledgments

Preparation of this article was supported by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC), the Ontario Problem Gambling Centre, and the Nova Scotia Gaming Foundation.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173−1182.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley and Sons.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods Research*, *21*(2), 230−258.

Fabrigar, L. R., & Wegener, D. T. (2009). Structural equation modeling. In J. P. Stevens (Ed.), *Applied multivariate statistics for the social sciences* (pp. 537−595)., 5th ed. New York, NY: Routledge, Taylor & Francis Group.

Fabrigar, L. R., Wegener, D. T., & MacDonald, T. K. (2010). Distinguishing between prediction and influence: Multiple processes underlying attitude–behavior consistency. In C. R. Agnew, D. E. Carlston, W. G. Graziano, & J. R. Kelly (Eds.), *Then a miracle occurs: Focusing on behavior in social psychological theory and research* (pp. 162−185). New York, NY: Oxford University Press.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components. *Structural Equation Modeling*, *12*(3), 343−367.

Gorsuch, R. L. (1983). *Factor analysis*, 2nd ed. Hillsdale, NJ: Erlbaum.

Hu, L., & Bentler, P. M. (1998). Fit indexes in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424−453.

Iacobucci, Dawn (2009). Everything you always wanted to know about SEM (structural equations modeling) but were afraid to ask. *Journal of Consumer Psychology*, *19*, 673−680.

Iacobucci, Dawn (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*, 90−98.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19−35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130−149.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84−99.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320−341.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York, NY: Springer-Verlag.

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 41−72). New York, NY: Guilford Press.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231−251.

Wegener, D. T., & Fabrigar, L. R. (2000). Analysis and design for nonexperimental data: Addressing causal and noncausal hypotheses. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412−450). New York: Cambridge University Press.