International Journal of Behavioral Development
2007, 31 (4), 357–365
http://www.sagepublications.com

© 2007 The International Society for the
Study of Behavioural Development

DOI: 10.1177/0165025407077757

New developments in latent variable panel analyses of longitudinal data

Todd D. Little, Kristopher J. Preacher, and James P. Selig
University of Kansas, USA

Noel A. Card
University of Arizona, USA

We review fundamental issues in one traditional structural equation modeling (SEM) approach to analyzing longitudinal data – cross-lagged panel designs. We then discuss a number of new developments in SEM that are applicable to analyzing panel designs. These issues include setting appropriate scales for latent variables, specifying an appropriate null model, evaluating factorial invariance in an appropriate manner, and examining both direct and indirect (mediated), effects in ways better suited for panel designs. We supplement each topic with discussion intended to enhance conceptual and statistical understanding.

Keywords: longitudinal data; panel designs; SEM

Although numerous new methods for the analysis of longitudinal data have been developed in recent years, traditional panel designs remain popular and offer basic answers to questions that new methods do not directly address. The motivation for this article is to provide a detailed exploration of key issues that apply to panel designs. In addition, we incorporate recent advances in general applications of structural equation modeling (SEM). Although our discussion of these issues is conceptual in nature, real data examples are available at Quant.KU.edu.

This article addresses issues related to latent variable SEMs to longitudinal data. We do not elaborate on key issues such as the need for well-articulated theory, age-appropriate measures, or a requisite match between the statistical model and the theoretical questions (Collins, 2006; Embretson, 2007; Little, Bovaird, & Slegers, 2006). Instead, we discuss applications of traditional panel designs in contexts where the theory, measurement, and design issues are appropriately matched.

Longitudinal confirmatory factor analysis of panel data

The confirmatory factor analysis (CFA) model is a special case of SEM. For many applications, CFA represents the end point of one's analysis. For many other applications, however, CFA is the starting point for more elaborate model testing (Brown, 2006). In this regard, CFA is used to assess the adequacy of the measurement model so that the hypotheses pertaining to the structural relations among the constructs defined by the CFA can be tested with SEM. In longitudinal research, the CFA, or measurement, model answers basic questions about the nature of the constructs and the patterns of individual differences. Specifically, the longitudinal CFA addresses the questions: (1) Are the constructs measured equivalently across time? (2) Are the individual-differences standings in the constructs stable (or unstable) across time? (3) Are the within- and cross-occasion relations among the constructs stable or changing systematically over time (e.g., differentiation)? (4) Are the means and/or variances of the constructs stable or changing systematically over time?

In short, the longitudinal CFA addresses a number of validity-related issues, and because the constructs of interest are repeatedly assessed, these validity issues are more rigorously evaluated than is the case with cross-sectional data. Here, the content validity of each construct is addressed by examining the patterns and magnitudes of the factor loadings and intercepts as well as measurement equivalence over time. The criterion validity of each construct is addressed in a number of ways. First, the concurrent patterns of relations among the constructs within each time point are examined. This form of validity is generally established by evaluating whether the relations are consistent with an expected pattern and whether these concurrent relations replicate (if expected to do so) or change in an expected manner (e.g., differentiate, dedifferentiate) over time. Criterion validity is also addressed by examining the cross-time associations and determining if they conform to the expected a priori pattern (e.g., a simplex pattern). Although not usually described as criterion validity, expected similarities and differences in the mean structures of the constructs can also be evaluated to support the validity profile of the constructs. Finally, overall construct validity is evaluated deductively based on how well the content and criterion validities are supported by the data and the degree to which the patterns replicate the behavior of similar constructs in the literature.

Correspondence should be sent to Dr Todd Little, Department of Psychology, University of Kansas, 1415 Jayhawk Blvd., Lawrence, KS 66045–7555, USA; e-mail: yhat@ku.edu

We wish to thank Elizabeth McConnell for her assistance with some of the figures and Kevin Grimm for his comments and feedback. This study was supported in part by grants from the NIH to the University of Kansas through the Mental Retardation and Developmental Disabilities Research Center (5 P30 HD002528). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Clearly, CFA can be used to address important questions and provide a wealth of information about the sample and the constructs; however, to obtain these answers there are a number of modeling issues that should be considered and addressed.

Scale setting

The first issue that should be considered in fitting a longitudinal CFA concerns the scale(s) that should be used for the latent variables, a decision necessary in order to estimate the parameters of the model. This may seem trivial to many users of SEM, but in fact, scaling issues are critical (Blanton & Jaccard, 2006; Embretson, 2007; Gonzalez & Griffin, 2001; Little, Slegers, & Card, 2006). Users of SEM have overwhelmingly relied on the *marker variable* method of scaling constructs, perhaps by tradition or because many SEM software packages default to this approach. Briefly, the method consists of choosing one indicator of a construct and constraining its factor loading to 1.0 (and its intercept to 0 if mean structures are included). Not only does this serve as an identification condition (necessary to provide unique estimates of other model parameters), but it also equates the scale of the construct to that of the marker variable. As is well known, the choice of which marker variable to use is arbitrary, and the resulting scale of a given construct is, therefore, also arbitrary. That is, the scale of the construct is the scale of the arbitrarily chosen marker variable (Bollen, 1989). Similarly, fixing the latent variance to 1.0 and the latent mean to 0 (an approach that is often used in psychometrics research) loses any inherent meaning for the scale because the construct is scaled in a standardized (z-score) metric.

Little et al. (2006) introduced the *effects coding* identification method for situations in which the indicators of a construct are drawn from a pool of possible indicators that are measured on the same scale (i.e., essentially congeneric). This method is similar to the marker variable approach, but rather than fixing the loading and intercept of one indicator, one imposes the constraint that the average of the factor loadings is 1.0 and the average of the indicator intercepts is 0. When the scale mean and standard deviation of a construct have inherent meaning and would be of substantive interest, then the effects coding constraints provide a meaningful metric for the latent constructs. Estimating the parameters of constructs (variances, means, and covariances) in a meaningful metric provides estimates that are free of measurement error, readily interpretable, and generalizable/comparable across replications using the same constructs and indicators (Little et al., 2006).

Longitudinal factorial invariance

The factorial invariance literature is still quite active and various procedures and recommendations have been offered for establishing the tenability of invariance constraints (G.W. Cheung & Rensvold, 1999, 2002; Little, Card, Slegers, & Ledford, 2007; Vandenberg & Lance, 2000). Applying the logic of factorial invariance typically used in cross-group comparisons to the longitudinal case is relatively straightforward. The basic question here is: are the respective indicators representing the same underlying constructs over time? In longitudinal research, constructs can change in meaning or importance as one traverses different developmental epochs. Testing and establishing longitudinal factorial invariance provides empirical evidence that the fundamental meaning of the construct has not changed across the different developmental periods.

Factorial invariance is traditionally established by following a sequence of steps (or fitting a sequence of models). Specifically, one starts with an unconstrained model and progresses to more restricted (and nested) models to evaluate the tenability of each set of successively placed constraints (Little, 1997; Widaman & Reise, 1997). Figure 1 depicts a CFA for two constructs measured by three indicators each at three times of measurement.

The unconstrained model is commonly referred to as the configurally or form invariant model (Brown, 2006). Here, the pattern of indicator-to-construct relations is expected to be the same at each occasion. In this model, each construct has a scale-setting constraint (e.g., effects coded, marker variable) for both the mean and covariance structures, but no other constraints are placed on any of the parameters of the model. Importantly, in longitudinal panel models the residuals of corresponding indicators are also allowed to correlate across measurement occasions (see Figure 1). The reason for allowing the corresponding residuals to correlate over time is because residual information in each of the corresponding indicators has two sources of variability: a variance component that is reliable but specific to a given indicator, and a random, unreliable source. When an indicator is represented in a model at more than one time of measurement, the item-specific component would be expected to covary with itself across times of measurement.

The next level of measurement invariance is termed loading invariance or weak factorial invariance (Brown, 2006; Vandenberg & Lance, 2000; Widaman & Reise, 1997). This level of measurement invariance is established when the loadings of corresponding indicators are equated across time. If this level of invariance holds (i.e., if the constrained model fits well relative to a model without the equality constraints), any changes in the amount of reliable variance among the indicators is adequately captured as changes in the amount of common construct variance (i.e., this level of invariance does not assume that variances are equal over time; see Selig, Card, & Little, in press). In our discussion of model fit, we describe the various criteria by which the adequacy of the invariance constraints can be evaluated.

The third level of measurement invariance is termed intercept invariance or strong factorial invariance (Brown, 2006; Meredith, 1993; Vandenberg & Lance, 2000; Widaman & Reise, 1997). This level of measurement invariance is established when, in addition to the loadings, the intercepts of corresponding indicators are equated across time. If this more stringent level of invariance holds, any changes in the mean levels of the indicators are adequately captured as changes in the underlying means of the latent constructs.

The fourth level of measurement invariance is termed residual invariance or strict factorial invariance (Brown, 2006; Meredith, 1993; Vandenberg & Lance, 2000; Widaman & Reise, 1997). This level of measurement invariance is established when the residual variances of corresponding indicators are equated across time. If this level of invariance holds, then the sum of the item-specific and random sources of measurement error variance for each indicator does not change over time. This level of invariance is rarely enforced in practice because it reflects a level of restriction that is usually

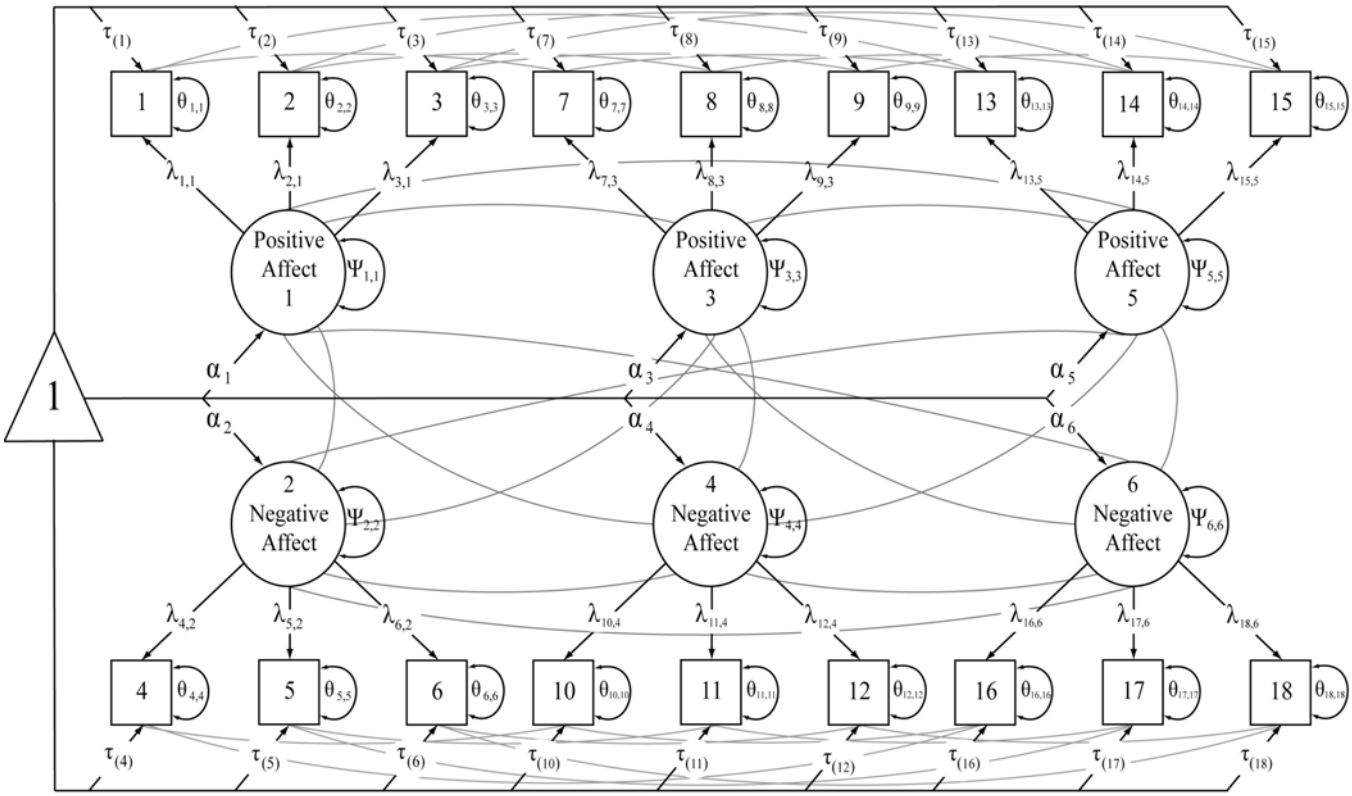

Figure 1. A detailed schematic of a longitudinal CFA. Not all parameters presented in this figure are identified unless scaling and identification constraints are place on various elements. See text for a description of three different methods of identification for such a model.

unrealistic to expect (Brown, 2006; Little, 1997; Widaman & Reise, 1997).

Evaluating invariance constraints

Factorial invariance across at least two groups or occasions is traditionally tested by specifying a series of models, each more constrained than the last, representing various degrees of invariance. Equality constraints are placed on key parameters, and the resulting change in model fit (usually in terms of the change in χ^2 relative to the change in degrees of freedom (*df*), respectively $\Delta\chi^2$ and Δdf) is monitored to assess whether adding the constraint was justifiable (Vandenberg & Lance, 2000; Widaman & Reise, 1997). Only after invariance has been satisfactorily demonstrated can panel models with latent variables be interpreted (at a minimum one should have at least partial weak invariance to examine covariance relations and partial strong invariance to examine mean structures).

In longitudinal designs, a first step is to test whether the occasion-specific covariance matrices and mean structures are equal across waves of measurement. If they are equal, there is little need to proceed with further invariance testing because this indicates no differences in the covariance and mean structure across time. If they are not equal, the researcher might proceed to apply equality constraints consistent with the levels of invariance described earlier until the most appropriate level has been identified. The utility of proceeding to a more strict level of invariance is traditionally gauged via a $\Delta\chi^2$ test with as many *df* as there were constraints added (cf. Little, 1997, on using a modeling rationale).

There are limitations inherent in this method, however (Preacher, Cai, & MacCallum, 2007). First, as with any null hypothesis test, we know the test to be false before we even collect data. The difference in the fit of two nested models is never literally 0, so the hypothesis being tested could be considered either absurd or uninteresting. Second, the $\Delta\chi^2$ test is known to be very sensitive to sample size; larger N values necessarily result in larger values of $\Delta\chi^2$, so the nested model strategy is biased in favor of invariance when N is small and is biased against invariance when N is large (MacCallum, Browne, & Cai, 2005).

Nested model tests for invariance do not represent the only viable approach for identifying the level of invariance characterizing a data set. Determining the level of invariance that best characterizes a data set may depend in part on the goals of the modeler. If the researcher's goal is to identify the true level of invariance, there must exist a 'true level' to identify. Strictly speaking, however, *no* level of invariance is correct for the simple reason that no model is correct, even when fit is perfect. Models are simply convenient mathematical or schematic representations of theoretical predictions, and are not expected to directly map onto the processes they represent. It is more realistic to say that, although no model is ever strictly correct, many models may provide useful approximations to the truth (Cudeck & Henly, 1991; MacCallum, 2003). If the researcher's goal is to find the best approximation (from the pool of available alternative models) to the data-generating process, then fit indices designed to reflect population-level approximation error, such as the RMSEA, should perform the best.

Good fit is only one criterion that has been suggested to characterize good models. Replicability is another. Therefore,

an alternative strategy is to identify the model (level of invariance, in this context) that best replicates when fit to another sample from the same population. Replication is the principle behind cross-validation. If identifying a model that maximizes replicability is the researcher's goal, then it is sensible to use model selection criteria that were designed with this goal in mind, such as the expected cross-validation index (ECVI; Browne & Cudeck, 1989, 1993), AIC, or BIC.

The bottom line is that when assessing factorial invariance – either cross-group invariance or longitudinal invariance with panel data – it is important to clearly articulate the goal. If the goal is to identify the best-fitting level of invariance for the data in hand, then absolute fit indices such as RMSEA may be more appropriate. If the goal is to identify the level of invariance most likely to replicate in future samples, then selection criteria may be preferable.

Specifying an appropriate null model

Widaman and Thompson (2003) discuss a number of issues regarding relative fit indices. Relative (or incremental) fit indices are routinely used to evaluate the overall fit of a model. Commonly used fit statistics include the comparative fit index (CFI) and the non-normed fit index (NNFI), which is the same as the Tucker–Lewis Index (TLI) commonly used in exploratory factor analysis. All of the relative fit statistics have in common a reliance on a null model, a worst-fitting model nested within the hypothesized model. Most SEM software packages fit a null model and calculate the relative fit statistics automatically. By default, the null model is usually the *independence model*. The independence model assumes zero covariance among the indicators (variables) in a model, but freely estimates the variances of these indicators. For many applications (e.g., single-group, single-occasion models), the independence model is a reasonable null model because it helps calibrate how much covariation among the indicators can be recovered by a substantively meaningful model.

For longitudinal (and multiple group) models, the independence assumption is only one piece of the 'null' expectation. When a researcher is interested in evaluating whether means and/or variances change across time, the independence model (in which one equates means or variances across time to evaluate this contribution to misfit) is not nested within such models, so the independence model no longer represents an adequate null model. Instead, the appropriate null model is one in which neither the variances nor the means of corresponding indicators change over time (Little, Card, Slegers et al., 2007; Widaman & Thompson, 2003). Because SEM software currently does not have the option to specify an alternative null model, users must specify and estimate a null model and use the fit information in the formula for calculating the various relative fit indices (calculators and sample code for specifying an appropriate null model are available at Quant.KU.edu).

Longitudinal structural models of panel data

We now turn our attention to key issues in fitting structural models to longitudinal panel data. Building upon the best fitting measurement model (assuming that at least partial weak invariance has been established), structural models are designed to address questions about directional patterns of effects. Specifically, structural models attempt to answer key questions about the pattern of direct (both autoregressive and cross-lagged) and indirect (i.e., mediated) relations among the constructs over time.

Causality in panel designs

Because of the temporal arrangement of constructs in longitudinal panel designs, the temptation to infer causal relationships in the patterns of direct and indirect influences is strong. Unfortunately, causality can be implied only in longitudinal panel data. Key threats to causal inferences include the exogeneity assumption, the omitted variable problem, and the potential confound of instrumental variables (i.e., covariates).

In common SEM parlance, latent variables that are represented in a model with no directed paths predicting them are termed exogenous variables, whereas variables that are predicted are termed endogenous variables. The exogeneity assumption refers to the idea that a longitudinal study begins at the start of a causal chain of events; however, the variables that are exogenous in the model might not be the true originating (causal) part of the system of changes. Even though exogeneity is an assumption of these models, the true causal agent in the system of change being modeled may have occurred at an earlier point in time. Violating this assumption may be less problematic if one is focused on proximal causes, but may still be relevant if a second proximal cause precedes a modeled proximal cause.

The omitted variable problem is similar to the exogeneity assumption, in that one might wrongly presume to have modeled the true causal constructs that are part of the system under scrutiny. Here, the causal agent in the system may be a common, but unmeasured or omitted, variable that causes two or more of the variables being modeled to covary, giving the appearance of causality. In some senses, the exogeneity assumption can be seen as a special case of the omitted variable problem.

A related problem is the proxy variable problem. The proxy variable problem describes a pervasive threat to validity – namely, that the measured construct may be only a proxy of the intended construct. For detailed examples of the proxy variable problem see Little, Card, Bovaird, Preacher, and Crandall (2007).

Including covariates

Potential covariates are commonly considered, and sometimes are even included in a particular analysis. The proper use of covariates, however, is not widely understood. For example, one might include gender as a covariate; however, the method for including such a variable is subject to choice. In a longitudinal panel study, such a covariate can be included in at least four different ways. First, the effects of gender can be partialed from all the indicators of all the constructs. This approach would involve specifying a model in which each indicator is regressed on the gender construct, thereby removing any variance shared with gender from each indicator. This approach can be used even prior to the construction of an SEM model by calculating the covariances among the indicators patialed for gender and using these partialed sufficient statistics as input for an SEM package. This latter approach is not recommended because the covariate effects are not explicitly represented in a model (and, therefore, issues of effect size, significance, overall model fit, and degrees of freedom related

to the covariate are not evaluated) but may be appropriate under some circumstances (e.g., when model complexity becomes unwieldy or when the significance and magnitude of covariate effects are not of concern).

A second approach to representing the effects of a covariate (such as gender) is to model the covariate as a direct effect on each latent construct, regardless of time of measurement. This approach is similar to modeling partialed indicators, but now the partialing is established for the latent variables instead of the manifest variables. In situations where the covariate is a discrete variable such as gender, multiple-group models can be specified to examine the homogeneity across the groups to see if it is warranted to collapse the covariance information across groups and represent the mean differences of gender as a covariate in a model only when the data are collapsed across the groups. If the homogeneity of the variance–covariance matrices across groups is not tenable, then the covariate (e.g., gender) has a moderating effect on the relations among the variables and the analyses should be performed in the multiple-group format. If homogeneity of the variance–covariance matrix is tenable and mean differences exist, data can be collapsed, and including the covariate as a variable in the model would adequately control for the mean differences.

A third approach is to control for covariate differences as effects on the constructs at the first measurement occasion, thereby accounting for the influence of the covariate 'downstream' in the model via various indirect pathways. This approach assumes that the covariate influence is an exogenous process and that once this effect is accounted for at the first occasion of measurement, the influence of the covariate no longer has a direct impact on the endogenous constructs.

As a fourth approach, one can control for the covariate effect at the final occasion of measurement. This approach is similar to entering a covariate first in a regression equation to remove its influence from the dependent variable before entering other predictors.

Deciding on which approach to use depends on the nature of covariates included and the goals of the modeling endeavor. Moreover, further work is needed to articulate the substantive implications of each approach and the conditions under which one approach would be more appropriate than another.

Mediation in panel designs

Mediation is the process by which an intervening *mediator* variable carries the effect of an independent variable (X) to a dependent variable (Y). In other words, M is the mechanism by which X exerts its effect on Y. For example, parental involvement in school influences a child's motivation for school, which in turn influences the child's actual school performance. There is a rich methodological literature on methods for assessing the magnitude and significance of mediation effects (or indirect effects; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Shrout & Bolger, 2002). Most modern approaches quantify the indirect effect as the product of the path coefficient linking X to M (denoted *a*) and the coefficient linking M to Y while controlling for X (denoted *b*). The various methods differ mainly in how the significance of this quantity is assessed. In panel models, there is a further complication in that there may be many *a* and *b* paths from which to choose. In what follows we first say a few words about causal inference. We follow this with a discussion of how the statistical significance of mediation effects may be ascertained.

Causal inference

A pervasive flaw in many studies addressing mediation effects concerns the fact that they are often assessed using cross-sectional data. Mediation is a causal process, however, and a fundamental prerequisite for making claims of causality is temporal separation. That is, measurement of variables involved in causal processes must be separated by enough time to permit the causal effect to unfold. Cole and Maxwell (2003) point out that mediation hypotheses tested against cross-sectional data (i.e., the majority of those seen in the literature) can be biased and very misleading. Panel designs are ideally suited for correcting this common problem.

Cole and Maxwell (2003) suggest that a model like that in Figure 2 be employed, in which X, M, and Y are each measured at several occasions in a panel design. Making the assumption of stationarity (i.e., that the causal effects do not change in magnitude over time) permits the researcher to equate the *a* paths to equality across occasions, as well as the *x*, *m*, *y*, and *b* paths (the tenability of these constraints is testable via nested model comparisons). Omitted from the diagram, but present in the model, are c' paths linking X measured at time $t-2$ to Y measured at time t. Also omitted are within-occasion residual covariances, which may be included or not, as the situation decrees. Use of a model like that in Figure 2 is far preferable to the traditional cross-sectional analysis of mediation. First, the temporal lag necessary for causal inference is explicitly considered. Second, repeated measurement of the key variables permits more accurate estimation of the path coefficients than in a cross-sectional design.

However, several issues need to be considered before such a model is used in assessing mediation (Cole & Maxwell, 2003). First, the optimal lag separating measurement waves needs to be identified. It is likely that the magnitude of the path coefficients will shift as the lag becomes shorter or longer, so pilot research is sometimes necessary to identify theoretically appropriate lags. Second, it is critical that unreliability be minimized. Unreliable variables can either attenuate or spuriously inflate path coefficients, usually in ways impossible to predict beforehand. Fortunately, reducing the biasing effects of unreliability is one of the primary benefits of using latent variable models.

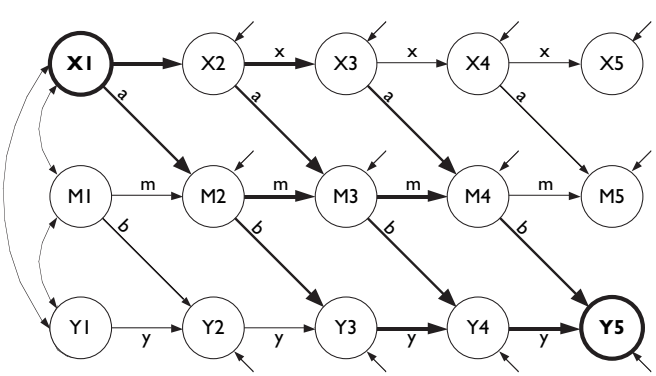

Figure 2. Type of panel model Cole and Maxwell (2003) suggest for testing mediation hypotheses linking X1 to Y5. All path coefficients involved in connecting X1 and Y5 are involved in quantifying the indirect effect.

Finally, estimation and testing of the indirect effect becomes more difficult in lagged panel designs, as we describe later. See Cole and Maxwell (2003) and Gollob and Reichardt (1985, 1987, 1991) for more detail.

Cole and Maxwell (2003) note that the longitudinal data collection needed to use the panel strategy for assessing mediation can be conveniently shortened by using only two waves of data collection. For example, in Figure 2 the paths linking X to M are constrained to equality over time. This assumption of stationarity is usually directly testable by merely setting the constraint and noting any significant decrement in fit. In addition, it is presumed that the optimal lags between measurement of X and M, and between measurements of M and Y, have been identified and used. If only two waves of data are collected, the model in Figure 3 may be employed, and mediation assessed as previously described. Here, stationarity must be assumed true, either on the strength of theory or prior research. Furthermore, the optimal lag between measurement of X and M must be assumed equal to the optimal lag between measurement of M and Y. Besides curtailing the cost of a study in terms of both time and money, employing a two-wave design has an added potential benefit in that it permits the researcher to experimentally manipulate M to strengthen causal inferences with regard to the M → Y effect. However, it is unclear whether the product of *a* and *b* will be meaningful in these circumstances because individual differences in M may have different meaning when M is manipulated versus observed.

Assessing statistical significance

Assessing the statistical significance of mediation effects is a very active area of research. Four general methods of assessing mediation effects prevail in the literature: the *causal steps* approach, the *product of coefficients* approach, the *distribution of the product* approach, and the *resampling* or *bootstrapping* approach. Other strategies have been suggested. Readers in need of a clear, basic introduction to mediation should consult Frazier, Tix, and Barron's (2004) excellent and approachable article describing and distinguishing mediation and moderation.

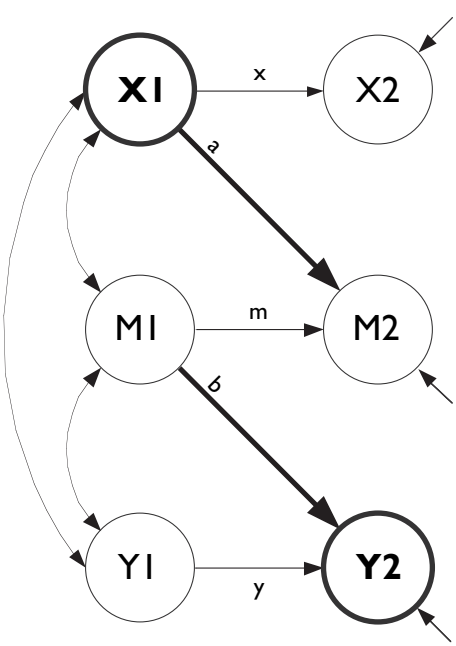

Figure 3. A two-wave panel model for testing mediation hypotheses.

Most social scientists associate tests for mediation with the *causal steps* criteria described by Baron and Kenny (1986). These criteria are as follows. First, X must significantly predict Y. This *total effect*, quantified as the path coefficient linking X and Y, is sometimes designated *c*. Second, X must predict M (path *a*). Third, M must significantly predict Y, conditional on the presence of X in the model (path *b*). Fourth, upon addition of M as a mediator, the *direct effect* of X on Y (c') should decrease relative to *c*. This pattern of effects is consistent with the notion that X explains variability in Y because X predicts M, which in turn predicts Y. Satisfying these criteria remains the most popular approach to gauging the extent and significance of mediation effects. However, several drawbacks are associated with it. The causal steps approach suffers from lower power than alternative methods (MacKinnon et al., 2002) and does not address the hypothesis of mediation directly. Rather, the researcher must infer the presence or absence of mediation by interpreting the pattern of relevant regression weights. Furthermore, in complex panel designs like that in Figure 2, there are no longer straightforward *a*, *b*, and c' paths to map onto the causal steps framework.

Whereas the indirect effect in simple cross-sectional models is quantified as *ab*, the *overall indirect effect* linking X at the first occasion to Y at the last occasion in Figure 2 is $aby^2 + amby + am^2b + xaby + xamb + x^2ab$. We will refer to sample estimates of complex indirect effects like this as $\hat{\omega}$. Assessing the significance or precision of such effects is considerably more difficult than in the traditional three-variable case. As with any sample statistic, it would be useful to know something about the sampling distribution of $\hat{\omega}$. In particular, being able to make the assumption of normality for $\hat{\omega}$ across repeated sampling would allow us to gauge its significance with a simple *z*-test, or (preferably) to use the point estimate $\hat{\omega}$ in a 95% confidence interval to estimate the population ω.

If the assumption of normality is appropriate, the problem is an easy one. Assuming that all the relevant path coefficients are normally distributed across repeated sampling (usually a safe assumption in large samples and with well-behaved data) means that the standard error (*SE*) of $\hat{\omega}$ can be derived using widely known methods. Various forms of this *SE* have been presented in connection with large-sample tests in simple mediation models (e.g., Sobel's test for mediation; Sobel, 1982; MacKinnon, Warsi, & Dwyer, 1995). Essentially a *z*-test of $\hat{\omega}$ when only one indirect path connects X and Y, the Sobel test is included in SEM software, is available online and in macros for popular statistics packages, and has become popular lately in the applied literature as an adjunct to reporting the results of the causal steps approach.

Unfortunately, the form of the sampling distribution of $\hat{\omega}$ – or, more precisely, the set of assumptions that may legitimately be made about the distribution – has been the subject of much debate. Early statistics literature makes it clear that $\hat{\omega}$ is not normally distributed even in simple models, although it may approach normality in large samples (Aroian, 1947; Craig, 1936). For most samples used in psychology, the distribution of $\hat{\omega}$ will be skewed and kurtotic, implying that the assumption of normality may not be tenable. Furthermore, the standard error of $\hat{\omega}$ increases greatly in complexity with the size of the model.

The most accurate and powerful approaches to assessing

indirect effects fall into two categories: the *distribution of the product* approach and *resampling* approaches. The distribution of the product approach accurately considers the correct distribution of $\hat{\omega}$ rather than simply assuming normality. Unfortunately, the distributional theory for $\hat{\omega}$ has been worked out only in the special case of a simple mediation model, and has not been extended to more complex panel designs.

Resampling, or *bootstrapping* (Bollen & Stine, 1990; Efron & Tibshirani, 1998; Lockwood & MacKinnon, 1998; MacKinnon, Lockwood, & Williams, 2004; Preacher & Hayes, 2004; Shrout & Bolger, 2002), recreates the sampling distribution of $\hat{\omega}$ by drawing thousands of *resamples* of size N (with replacement) from the original sample of size N, computing sample estimates of all relevant path coefficients, and forming a distribution of $\hat{\omega}$. The researcher may form a percentile-based confidence interval of any degree of precision α by locating the values of $\hat{\omega}$ cutting off the lower and upper (50α) % of the bootstrapped distribution. The results of any point hypothesis test are implied by the resulting interval. A modified form of bootstrapping, in which the confidence limits are adjusted slightly to correct for bias, has been found to perform very well in terms of type I error rates and statistical power (MacKinnon et al., 2004). The primary benefits of bootstrapping are that it involves no distributional assumptions, is characterized by highly accurate type I error rates and relatively high statistical power, and can be used in small samples. Resampling methods are available in some SEM software applications (M.W.-L. Cheung, in press; Shrout & Bolger, 2002).

Finally, we note a new and promising method of assessing the significance of mediation effects described by M.W.-L. Cheung (in press). M.W.-L. Cheung's method involves defining a new parameter and setting it equal to the expression for ω. A *likelihood-based confidence interval* for ω may be obtained by iteratively determining the values that $\hat{\omega}$ must assume for the overall fit of the model to change by $\pm.84$ units (a significant χ^2 when $df = 1$ and $\alpha = .05$). Currently only the SEM package Mx automatically computes likelihood-based intervals. The method is sufficiently general to compute intervals for indirect effects of any degree of complexity with equal ease. In panel designs like those in Figure 2, we recommend creating either bootstrapped confidence intervals (easiest in Mplus) or likelihood-based confidence intervals for ω (for straightforward descriptions of both methods, see M.W.-L. Cheung, in press).

Regardless of what strategy is used to assess mediation, it is strongly recommended that emphasis be placed on estimating the magnitude of the effect with confidence intervals rather than on making a dichotomous accept–reject decision (Wilkinson & the Task Force on Statistical Inference, 1999). There are many arguments favoring this emphasis. First, researchers already know the answer to the question 'is the null hypothesis (of no mediation) true?' The answer is invariably 'no' – it is a matter of direction and degree. Second, point (nil) hypothesis testing has become something of a mindless ritual (Gigerenzer, 2004), encouraging scientists to pursue low p-values rather than estimate the size of an effect in the population. Third, whereas comparing a p-value to α leads only to a dichotomous decision about the null hypothesis, confidence intervals can be used to infer the results of any hypothesis test, and add information about the magnitude of the effect and precision of estimation. In other words, emphasis should be placed on estimating how large an effect *is*, not how small it *isn't*.

Accelerated designs

It is always advantageous to collect data over many occasions when evaluating developmental processes. For example, the magnitude the cross-time correlation for a given construct might be expected to change with age, or one might expect that cross-lag paths differ across development. In these situations, researchers might consider collecting longitudinal data over a wide time frame (e.g., yearly assessment over the course of several years), usually at great cost in terms of money and time. Alternatively, researchers might consider accelerated longitudinal designs as a way of approximating these long-term longitudinal data.

Accelerated longitudinal designs, first proposed by Bell (1953) as an efficient method of obtaining longitudinal data over an extended developmental period, consist of shorter term longitudinal studies of several cohorts, linked so as to provide comprehensive coverage of a particular developmental period. Although these designs are more often discussed in the context of growth curve modeling (Meredith & Tisak, 1990; Tisak & Meredith, 1990), they also have value in fitting panel models. Figure 4 provides an illustration of an accelerated longitudinal path model, in which data are collected longitudinally for 2 or 3 years (more on this point below) from three cohorts of children aged 11, 12, and 13 at the initial assessment. An important requirement of this design is that there exist one or more points of linkage – ages where multiple cohorts are measured[1] (e.g., the youngest cohort was measured at time 2 when they were 12 years old, and the middle cohort was measured at time 1 when they were 12 years old).

Although these accelerated longitudinal designs are often treated as analyses with large amounts of missing data (Allison, 1987; McArdle & Hamagami, 1991), there are also advantages to treating them as multiple group SEMs. One advantage of the multiple group approach is that it allows for tests of measurement invariance not only across time (see above), but also across cohorts, thus evaluating whether the constructs are measured equivalently across time and cohort. A second advantage is that the process of evaluating cohort and developmental similarities and differences can be performed quite simply as nested model comparisons. This second advantage will be elaborated next.

Consider first the situation in which there is minimal overlap among cohorts, where there is only one shared age of measurement for each pair of successive cohorts. This is represented by the portions of the model with solid lines in Figure 4. Within a three-group model, one begins by evaluating measurement invariance across time as described above, then evaluating whether it is also tenable to equate loadings and intercepts across groups (i.e., cohorts; alternatively, one may choose to establish invariance across time and cohort simultaneously). Assuming measurement invariance is tenable, one then estimates the autoregressive (i.e., stability) and cross-lagged paths

[1] This requirement is, technically speaking, overly restrictive. Although most accelerated longitudinal designs follow this pattern, one could instead design a study in which the ranges of ages, rather than specific ages at assessment, overlapped. For instance, for this hypothetical example, one could assess the middle cohort at ages 11.5, 13, and 14.5 in order to obtain an overlap in age ranges rather than specific ages. It is not clear the extent of settings in which such a design would be advantageous; readers should be aware that data from these unbalanced accelerated longitudinal designs could be analyzed.

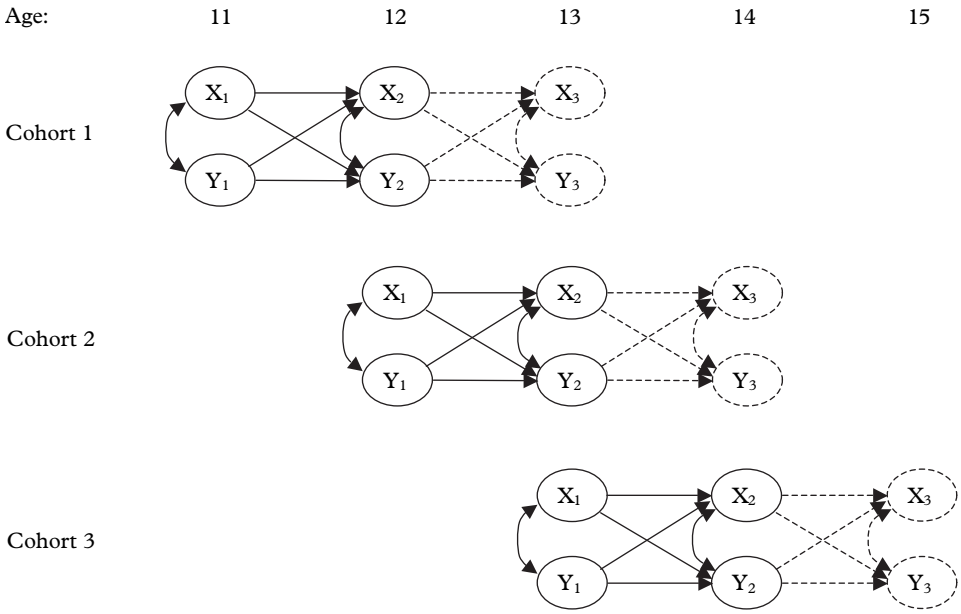

Figure 4. An example of an accelerated longitudinal panel design.

of interest, allowing these to be freely estimated across groups. From this model, one can then constrain the set of paths of interest (e.g., X predicting Y) to be equal across the three cohort groups. If this constraint does not result in a significant decrement in model fit (evaluated perhaps by $\Delta\chi^2$ on Δdf), then one can conclude that this process operates similarly across this developmental period. If this change is significant, however, this indicates that the process operates differently across cohort or development. In this minimal overlap condition, however, it is not possible to determine whether this change is due to developmental or cohort differences.

A second situation occurs when the researcher has two or more points of linkage between successive cohorts, as displayed in Figure 4 by the inclusion of portions with dashed lines. As in the minimal overlap situation, one begins by establishing measurement invariance across time and cohort, fitting an unrestricted model in which the paths of interest are freely estimated across time and cohort, then evaluating whether a model in which the paths of interest are constrained equal across time and cohort leads to a significant drop in model fit. If this test does indicate a significant drop, one can now evaluate the extent to which this is due to cohort versus developmental effects. To evaluate cohort effects, one would evaluate the increase in χ^2, relative to the unrestricted model, that occurs with the equating of paths of interest across cohorts during developmental periods where the cohorts overlap. For example, one might constrain the cross-lag path of X predicting subsequent Y to be equal for the following overlaps: (1) time 2 to time 3 for cohort 1 = time 1 to time 2 for cohort 2; and (2) time 2 to time 3 for cohort 2 = time 1 to time 2 for cohort 3. The decrement in model fit resulting from these constraints would indicate the effects of cohort differences. The remaining decrement in model fit from the unrestricted model to the fully restricted model (i.e., that where the parameter of interest is equated across time and cohort) could be interpreted as a developmental effect.

Conclusions

Traditional panel designs are still developing in terms of what would be considered best-practice approaches (e.g., specifying an appropriate null model) and in terms of choices that the investigator can make (e.g., scaling choices, model evaluation, how to include covariate effects, and how best to assess mediation). Moreover, data collected for traditional panel designs often lend themselves to more advanced accelerated designs. We feel that the traditional panel design still has much to offer the substantive researcher because it addresses specific questions about the nature of change that are not necessarily contained in other advance approaches such as growth curve modeling. In addition, recent advances in general SEM practices serve to augment the validity and utility of the information that is derived from such models. In sum, we encourage the continued use of panel designs when the research questions match their empirical content. By outlining recent advances in the application of SEM, we hope the quality of research using these designs will continue to advance.

References

Allison, P.D. (1987). Estimation of linear models with incomplete data. In C.C. Clogg (Ed.), *Sociological methodology* (pp. 71–103). San Francisco: Jossey-Bass.

Aroian, L.A. (1947). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics*, *18*, 265–271.

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

Bell, R.Q. (1953). Convergence: An accelerated longitudinal approach. *Child Development*, *24*, 145–152.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27–41.

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

Bollen, K.A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Browne, M.W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*, 445–455.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cheung, G.W., & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27.

Cheung, G.W., & Rensvold, R.B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212.

Cheung, M.W.-L. (in press). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling*.

Cole, D.A., & Maxwell, S.E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558–577.

Collins, L.M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528.

Craig, C.C. (1936). On the frequency function of *xy*. *Annals of Mathematical Statistics*, *7*, 1–15.

Cudeck, R., & Henly, S.J. (1991). Model selection in covariance structures analysis and the 'problem' of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Efron, B., & Tibshirani, R.J. (1998). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Embretson, S.E. (2007). Impact of measurement scale in modeling developmental processes and ecological factors. In T.D. Little, J.A. Bovaird, & N.A. Card (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Erlbaum.

Frazier, P.A., Tix, A.P., & Barron, K.E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*, 115–134.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Gollob, H.F., & Reichardt, C.S. (1985). Building time lags into causal models of cross-sectional data. *Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 165–170). Washington, DC: American Statistical Association.

Gollob, H.F., & Reichardt, C.S. (1987). Taking account of time lags in causal models. *Child Development*, *58*, 80–92.

Gollob, H.F., & Reichardt, C.S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L.M. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*. Washington, DC: American Psychological Association.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every 'one' matters. *Psychological Methods*, *6*, 258–269.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Little, T.D., Bovaird, J.A., & Slegers, D. (2006). Methods for the analysis of change. In D. Mroczek & T.D. Little (Eds.), *The handbook of personality development* (pp. 181–211). Mahwah, NJ: Erlbaum.

Little, T.D., Card, N.A., Bovaird, J.A., Preacher, K., & Crandall, C.S. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T.D. Little, J.A. Bovaird, & N.A. Card (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Erlbaum.

Little, T.D., Card, N.A., Slegers, D., & Ledford, E. (2007). Representing contextual factors in multiple-group MACS models. In T.D. Little, J.A. Bovaird, & N.A. Card (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Erlbaum.

Little, T.D., Slegers, D.W., & Card, N.A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Lockwood, C., & MacKinnon, D.P. (1998). *Bootstrapping the standard error of the mediated effect*. Paper presented at the 23rd annual meeting of SAS Users Group International.

MacCallum, R.C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139.

MacCallum, R.C., Browne, M.W., & Cai, L. (2005). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19–35.

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., & Sheets, V. (2002). A comparison of methods to test the significance of the mediated effect. *Psychological Methods*, *7*, 83–104.

MacKinnon, D.P., Lockwood, C.M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128.

MacKinnon, D.P., Warsi, G., & Dwyer, J.H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41–62.

McArdle, J.J., & Hamagami, F. (1991). Modeling incomplete longitudinal data using latent growth structural equation models. In L. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change* (pp. 276–304). Washington, DC: American Psychological Association.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Preacher, K.J., Cai, L., & MacCallum, R. (2007). Issues in comparing covariance structures models. In T.D. Little, J.A. Bovaird, & N.A. Card (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Erlbaum.

Preacher, K.J., & Hayes, A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*, 717–731.

Selig, J.P., Card, N.A., & Little, T.D. (in press). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F.J.R. van de Vijver, D.A. van Hemert, & Y.H. Poortinga (Eds.), *Individuals and Cultures in Multilevel Analysis*.

Shrout, P.E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Sobel, M.E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco: Jossey-Bass.

Tisak, J., & Meredith, W. (1990). Descriptive and associative developmental models. In A. Von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 2, pp. 387–406). Boston: Academic Press.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Widaman, K.F., & Reise, S.P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K.J. Bryant, M. Windle, & S.G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychiatric Association.

Widaman, K.F., & Thompson, J.S. (2003). On specifying the null model for incremental fit indices in Structural Equation Modeling. *Psychological Methods*, *8*, 16–37.

Wilkinson, L., & and Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.