Journal of Applied Psychology
2005, Vol. 90, No. 4, 710–730

Copyright 2005 by the American Psychological Association
0021-9010/05/$12.00 DOI: 10.1037/0021-9010.90.4.710

The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions

Scott B. MacKenzie and Philip M. Podsakoff
Indiana University Bloomington

Cheryl Burke Jarvis
Arizona State University

The purpose of this study was to review the distinction between formative- and reflective-indicator measurement models, articulate a set of criteria for deciding whether measures are formative or reflective, illustrate some commonly researched constructs that have formative indicators, empirically test the effects of measurement model misspecification using a Monte Carlo simulation, and recommend new scale development procedures for latent constructs with formative indicators. Results of the Monte Carlo simulation indicated that measurement model misspecification can inflate unstandardized structural parameter estimates by as much as 400% or deflate them by as much as 80% and lead to Type I or Type II errors of inference, depending on whether the exogenous or the endogenous latent construct is misspecified. Implications of this research are discussed.

A substantial amount of attention has been paid in the past 25 years to the issue of construct validation in the behavioral and organizational sciences. Construct validation is important because, as Schwab (1980) has noted, establishing the substantive validity of a construct before examining its construct validity may lead to the accumulation of knowledge that later must be discarded: "Organizational behavior has suffered because investigators have not accorded construct validity the same deference as substantive validity. . . . As a consequence, substantive conclusions have been generated that may not be warranted" (p. 34).

Thus, it is not surprising that a considerable amount of effort has been devoted to developing procedures to improve the scale development process (cf. Hinkin, 1995; Nunnally & Bernstein, 1994; Schwab, 1980; Spector, 1992). These efforts are evident in the rise in the reporting of confirmatory factor analyses, convergent and discriminant validity, and internal consistency reliability as part of the scale validation process.

However, these procedures are all founded on classical test theory and its assumptions about the relationships between latent constructs and their measures. Classical test theory assumes that the variance in scores on a measure of a latent construct is a function of the true score plus error. Thus, meaning flows from the latent construct to the measures in the sense that each measure is viewed as an imperfect reflection of the underlying latent construct (cf. Bollen, 1989; Nunnally & Bernstein, 1994). For example, one could view a person's performance on a series of two-digit addition problems as a reflection of his or her "two-digit addition skill." Or one could view the following four items developed by Wong and Law (2002) as reflections of a person's ability to assess the emotions of others: "I always know my friends' emotions from their behavior," "I am a good observer of others' emotions," "I am sensitive to the feelings and emotions of others," and "I have good understanding of the emotions of people around me." The key point is that in this type of measurement model, the latent construct is empirically defined in terms of the common variance among the indicators.

Although this type of measurement model is conceptually appropriate in many instances, Bollen and Lennox (1991) have noted that it does not make sense for all constructs. Indeed, they argued that measures do not always reflect underlying latent constructs but sometimes combine to form them. This is consistent with the views of several other researchers (cf. Blalock, 1964; Bollen, 1984, 1989; Law & Wong, 1999; MacCallum & Browne, 1993) who have argued that for some latent constructs, it makes more sense to view meaning as emanating from the measures to the construct in a definitional sense rather than vice versa. For example, most researchers today conceptualize job satisfaction as comprising a variety of distinct facets, including satisfaction with one's work, pay, coworkers, supervisor, and promotion opportunities. From a conceptual perspective, these distinct facets of satisfaction together determine a person's overall level of job satisfaction. Thus, in this type of measurement model, the latent construct is empirically defined in terms of the total variance among its indicators, and the indicators only capture the entire conceptual domain as a group.

This is a critically important distinction, because many of the scale development procedures recommended in the literature only apply to constructs with reflective measures, and if they are applied to constructs with formative measures, they can undermine construct validity. For example, most texts on scale development processes (cf. Schwab, 1980; Spector, 1992) recommend that items that possess low item-to-total correlations should be dropped from a scale to enhance internal consistency reliability. Although this recommendation is appropriate in the case of reflective indicators, because the items are all sampled from the same content domain, if this recommendation is followed for constructs with formative

Scott B. MacKenzie, Department of Marketing, Kelley School of Business, Indiana University Bloomington; Philip M. Podsakoff, Department of Management, Kelley School of Business, Indiana University Bloomington; Cheryl Burke Jarvis, Department of Marketing, W. P. Carey School of Business, Arizona State University.

Correspondence concerning this article should be addressed to Scott B. MacKenzie, Department of Marketing, Kelley School of Business, Indiana University Bloomington, 1309 East 10th Street, Bloomington, IN 47405-1701. E-mail: mackenz@indiana.edu

indicators, it may result in the elimination of precisely those items that are most likely to alter the empirical and conceptual meaning of the construct. Thus, as noted by Bollen and Lennox (1991), the conventional wisdom on item selection and scale evaluation must be qualified by consideration of the directional relationship between the indicators and the latent construct.

The distinction between formative and reflective indicators is also important because failure to properly specify measurement relations can threaten the statistical conclusion validity of a study's findings. For example, Law and Wong (1999) have noted that measurement model misspecification can sometimes bias estimates of the structural relationships between constructs and potentially undermine statistical conclusion validity (although it did not do so in their study). If this were found to be generally true, it would suggest that measurement model misspecification may cause Type I and/or Type II errors of inference in hypothesis testing.

However, as yet it is not known just how much impact such misspecification might have or under what conditions it is likely to have biasing effects. In addition, little guidance exists for researchers about how to distinguish formative from reflective indicators or about how to develop, model, and evaluate constructs with formative indicators. Therefore, the purposes of this study were to (a) discuss the distinction between formative- and reflective-indicator measurement models, (b) develop criteria for deciding whether measures are formative or reflective, (c) illustrate constructs that should be modeled as having formative indicators, (d) empirically test the effects of measurement model misspecification using a Monte Carlo simulation, and (e) recommend new scale development and validation procedures for constructs with formative indicators.

Measurement Model Specification

As noted by Cook and Campbell (1979), Nunnally and Bernstein (1994), Schwab (1980), and others, researchers use multiple measures of their constructs because (a) most constructs cannot be measured without error, (b) it is difficult for a single indicator to adequately capture the breadth of a construct's domain, and (c) it is necessary to unconfound the method of measurement from the construct of interest. Thus, the use of multiple measures with maximally different methods is the best way to ensure that the measures validly and reliably represent the construct of interest. However, once a researcher has developed multiple measures, he or she faces the problem of how to accurately model the relationships between the measures and the construct of interest. Generally speaking, two different measurement models have been mentioned in the structural equation modeling literature: the common latent construct model with reflective indicators and the composite latent construct model with formative indicators.

Common Latent Construct Model With Reflective Indicators

Models of this type posit that covariation among measures is explained by variation in an underlying common latent factor. It is for this reason that the indicators are referred to as *effects* indicators (Bollen, 1989; Bollen & Lennox, 1991; MacCallum & Browne, 1993) that are *reflective* of the underlying construct they represent. This is illustrated in Figure 1A by an ellipse with several arrows emanating from it to a set of indicators. We refer to the factors in this model as *common* latent constructs for two reasons. First, this is the most common type of measurement model found in the behavioral and organizational literature. Second, the latent construct is empirically defined in terms of the common (shared) variance among the items.

As noted by Bollen and Lennox (1991), there are several key features of this type of measurement model that should be recognized. First, the direction of causality flows from the construct to the measures in the sense that the construct explains the variation in the measures. Second, the indicators in this type of measurement model should be highly correlated due to the fact they all reflect the same underlying construct. As a result, they should exhibit high levels of internal consistency reliability. Third, "for all practical purposes, equally reliable effect indicators of a *unidimensional* [construct] are interchangeable" (Bollen & Lennox, 1991, p. 308). This is true because each of the measures is supposed to be sampled from the same conceptual domain and to represent all aspects of it. This implies that dropping one of two equally reliable indicators from the measurement model should not alter the meaning of the construct. Fourth, in this type of measurement model, error is associated with the individual measures rather than with the construct as a whole (though an overall calculation of the reliability of a set of measures can be made on the basis of the individual measure reliabilities). One advantage of this is that it

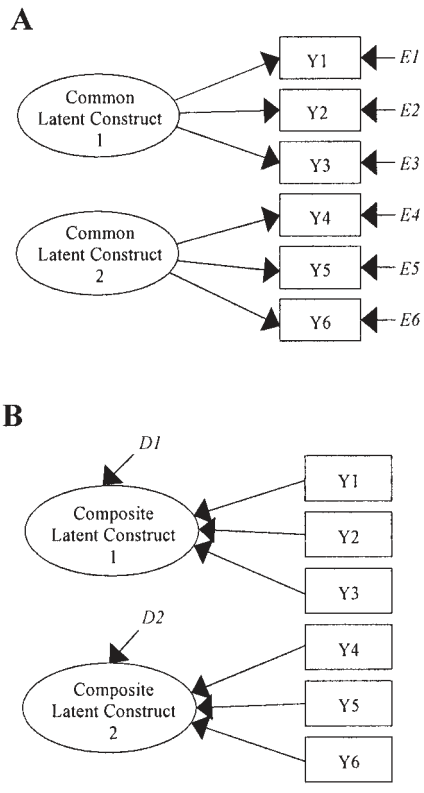

Figure 1. Factor specification for the common latent construct model with reflective indicators (A) and the composite latent construct model with formative indicators (B).

permits researchers to evaluate the differential reliability of the individual items in their scales. This is helpful when designing scales because it provides a basis for identifying weaker items and suggests areas where the scale could be improved. Finally, because the measures are all imperfect reflections of the underlying construct, a summed scale score will not adequately represent a construct with reflective indicators, and using a scale score in place of the latent construct will result in inconsistent structural estimates of the relationships between the construct and other latent constructs.

Composite Latent Construct Model With Formative Indicators

Less well known than the reflective-indicator measurement model is the *formative*-indicator (or *causal*-indicator) measurement model. As indicated in Figure 1B, this model posits that the measures jointly influence the composite latent construct, and meaning emanates from the measures to the construct in the sense that the full meaning of the composite latent construct is derived from its measures. This has two important implications. First, because the measures are not hypothesized to be caused—or determined—by the composite latent variable, the model itself does not assume or require the measures to be correlated (cf. Bollen, 1984; Bollen & Lennox, 1991). Indeed, it would be entirely consistent with this measurement model for the indicators to be completely uncorrelated. Therefore, internal consistency reliability is not an appropriate standard for evaluating the adequacy of the measures in formative models. Indeed, as noted by Bollen and Lennox (1991), "causal [formative] indicators are not invalidated by low internal consistency so to assess validity we need to examine other variables that are effects of the latent construct" (p. 312). This would suggest that to assess the validity of formative indicators, researchers must pay particular attention to nomological and/or criterion-related validity.

A second implication is that the consequences of dropping a formative indicator from a measurement model are potentially much more damaging than the consequences of dropping a reflective indicator. Although dropping one of two equally reliable measures from a reflective-indicator model does not alter the empirical meaning of a construct, that may not be true for a formative-indicator model. Assuming the measures are not redundant (i.e., they tap different facets of the conceptual domain), dropping a measure from a formative-indicator model may omit a unique part of the conceptual domain and change the meaning of the variable, because the construct is a composite of all the indicators. This is true because unlike reflective measures that individually tap the entire conceptual domain, formative measures only capture the entire conceptual domain as a group. This suggests that for formative-indicator models, following the standard scale development procedures—that is, dropping the items that possess the lowest item-to-total correlations or the lowest factor loadings—may result in the removal of precisely those items that would most alter the empirical meaning of the composite latent construct. Doing so could make the measure deficient by restricting the domain of the construct (cf. Churchill, 1979; Schwab, 1980). Thus, this is another reason why measures of internal consistency reliability should not be used to evaluate the adequacy of formative-indicator models.

Related to the above discussion, because the composite latent variable is explained by the measures in a formative-indicator model, high intercorrelations between formative indicators can make it difficult to separate the distinct impact of the individual indicators on the construct. This happens because the indicator coefficients are analogous to those obtained from a multiple regression of the latent construct on the formative indicators, and the stability of these coefficients is influenced by both multicollinearity and sample size (cf. Bollen & Lennox, 1991). This makes it difficult to identify the unique effect of each indicator on the construct. This is not the case in the reflective-indicator model, in which the coefficients relating the latent construct to its indicators are analogous to simple regression coefficients. Thus, although multicollinearity may be viewed as a virtue for reflective indicators, it can be a significant problem for measurement-model parameter estimates when the indicators are formative.

A final feature of the formative-indicator model is that like the reflective-indicator model, it includes an error term. However, unlike the reflective-indicator model, error is represented at the construct level rather than at the individual-item level. The error estimate for this model captures the invalidity of the set of measures—caused by measurement error, interactions among the measures, and/or aspects of the construct domain not represented by the measures—rather than the amount of error attributable to each individual measure. The presence of a construct-level error term is also a reminder of the fact that a formative-indicator construct is more than just a shorthand way of referring to an empirical combination of measures. It possesses what MacCorquodale and Meehl (1948) termed *surplus meaning*:

> These constructs involve terms which are not wholly reducible to empirical terms; they refer to processes or entities that are not directly observed (although they need not be in principle unobservable); the mathematical expression of them cannot be formed simply by a suitable grouping of terms in a direct empirical equation; and the truth of the empirical laws involved is a necessary but not sufficient condition for the truth of these conceptions. (p. 104)

Consequently, as is true for reflective-indicator constructs, a construct with several formative indicators cannot be adequately represented by a summed scale score, and using a scale score to represent a formative-indicator construct will lead to biased estimates of the structural relationships involving the construct. As noted by Bollen and Lennox (1991),

> if the composite [scale score] is the only variable measured with error, then the coefficient estimated for that variable will tend to be too low. In the more realistic situations of more than one explanatory variable containing error, the coefficient estimates can tend to be downwardly or upwardly "biased." (p. 310)

This is true even if a weighted sum is used instead of an unweighted sum. The only time this would not be true is in the unlikely event that all of the coefficients relating the measures to the construct were equal to 1, and construct-level error was equal to 0.

Criteria for Distinguishing Between Reflective- and Formative-Indicator Models

Given the importance of the differences between formative and reflective measurement models, it is important for researchers to

carefully evaluate the nature of the relationships between their constructs and measures. The first question to consider is whether the indicators are defining characteristics of the construct or manifestations of it. If the measures represent defining characteristics that collectively explain the meaning of the construct, a formative-indicator measurement model should be specified. However, if the measures are manifestations of the construct in the sense that they are each determined by it, a reflective-indicator model is appropriate. This judgment can be made by carefully thinking about whether it is more likely that changes in the latent construct would produce changes in the measures than it is that changes in the measures would produce changes in the latent construct.

A second question is whether the indicators appear to be conceptually interchangeable. If the measures are reflective, they should share a strong common theme, and each of them should capture the essence of the domain of the construct. Indeed, reflective measures are typically viewed as being sampled from the same conceptual domain. However, this is not generally true for formative measures (cf. Bollen & Lennox, 1991). If the indicators are formative, they may not necessarily share a common theme, and each of them may capture a unique aspect of the conceptual domain.

Closely related to this, a third question to consider is whether the indicators would be expected to covary with each other. A reflective-indicator measurement model explicitly predicts that the measures should be strongly correlated with each other because they share a common cause (i.e., they all reflect the same underlying latent construct). In contrast, a formative-indicator measurement model makes no predictions about the correlations among the measures. They might be high, low, or somewhere in between. Thus, if the indicators are not expected to be highly correlated, a reflective-indicator measurement model would seem to be inappropriate. However, if the indicators are expected to be highly correlated, then either model might be appropriate, and one would need to rely on the other criteria.

A final question to consider is whether all of the indicators are expected to have the same antecedents and/or consequences. Reflective indicators of a construct should all have the same antecedents and consequences because they all reflect the same underlying construct and are supposed to be conceptually interchangeable. However, because formative indicators are not necessarily interchangeable and may tap unique aspects of the conceptual domain, they would not necessarily be expected to have similar antecedents and consequences. Therefore, if some of the measures are expected to have different antecedents and/or consequences, they should be modeled as formative indicators, whereas if they all share virtually the same antecedents and consequences, they should be modeled as reflective indicators.

A Continuum of Reflective- and Formative-Indicator Measurement Models

The distinction between reflective- and formative-indicator models can be generalized to higher order factor structures. Up to this point, the discussion has focused exclusively on the relationships between measures and first-order latent constructs. However, it is important to recognize that conceptual definitions of constructs are often specified at a more abstract, second-order level, with multiple first-order subdimensions serving as reflective or formative indicators (cf. Bacharach, Bamberger & Sonnenstuhl, 2002; Baum, Locke, & Smith, 2001; Demerouti, Bakker, Nachreiner, & Schaufeli, 2001; Holtom, Lee, & Tidd, 2002; Hom & Kinicki, 2001; Mitchell, Holtom, Lee, Sablynski, & Erez, 2001). Because of this, it is possible for a single multidimensional construct to have one type of measurement model relating its measures to its first-order subdimensions and a different measurement model relating its subdimensions to the second-order latent construct they represent. It is also possible for a construct to have a mixture of some reflective and some formative indicators at either level of abstraction.

Figure 2 illustrates a series of three related models with reflective indicators, and Figure 3 depicts a corresponding series of models with a mixture of reflective and formative indicators. The first panel in each figure represents, respectively, the common latent construct and composite latent construct models discussed above. The other panels represent important elaborations of these initial models. There are several planes represented in each of these figures. The top one represents the conceptual plane, whereas the bottom one represents the observational plane. The middle planes represent first- and second-order empirical abstractions. The figures are drawn in this way to emphasize two important distinctions: (a) The latent constructs (whether first or second order) are empirical abstractions intended to formally represent the hypothetical constructs, but the two are not synonymous (cf. Bollen, 1989), and (b) neither the hypothetical constructs nor the latent constructs that represent them can be measured directly without error (cf. Bacharach, 1989; Schwab, 1980).

Figure 2 depicts the most commonly used measurement models in behavioral research. Panel 1 of this figure shows a unidimensional first-order latent construct with three reflective indicators. Panel 3 shows a series of first-order latent factors with reflective indicators, and it also shows that these first-order factors are themselves reflective indicators of an underlying second-order construct. Panel 2 also shows a second-order construct, with three first-order latent factors as reflective indicators. However, each of the facets in this panel has only a single reflective indicator. There are a couple of interesting features of these models that are important to recognize. First, the model shown in Panel 2 is not identified, because unique values for the item-level error terms and facet-level error terms cannot be simultaneously estimated. However, the model can be estimated if one or the other of these error terms is fixed (usually at the value of 0, although not necessarily so). Second, if the item-level error terms are fixed at 0, the models shown in Panels 1 and 2 are empirically indistinguishable. That is, their predicted covariance matrices are identical. Third, the models shown in Panels 2 and 3 are conceptually equivalent, differing only in the number of measures reflecting each of the facets. Generally speaking, the difference between these models is that the model in Panel 3 permits item-level measurement error to be distinguished from facet-level measurement error. This advantage is a direct result of the fact that multiple indicators of each facet are available. Finally, it is important to recognize that the models represent a continuum of conceptual differentiation in the sense that all three might apply to the same hypothetical construct. The choice would depend on the generality or specificity of one's theoretical interests.

To clarify this point, consider the case of a hypothetical construct such as "liking for a supervisor." One might define the

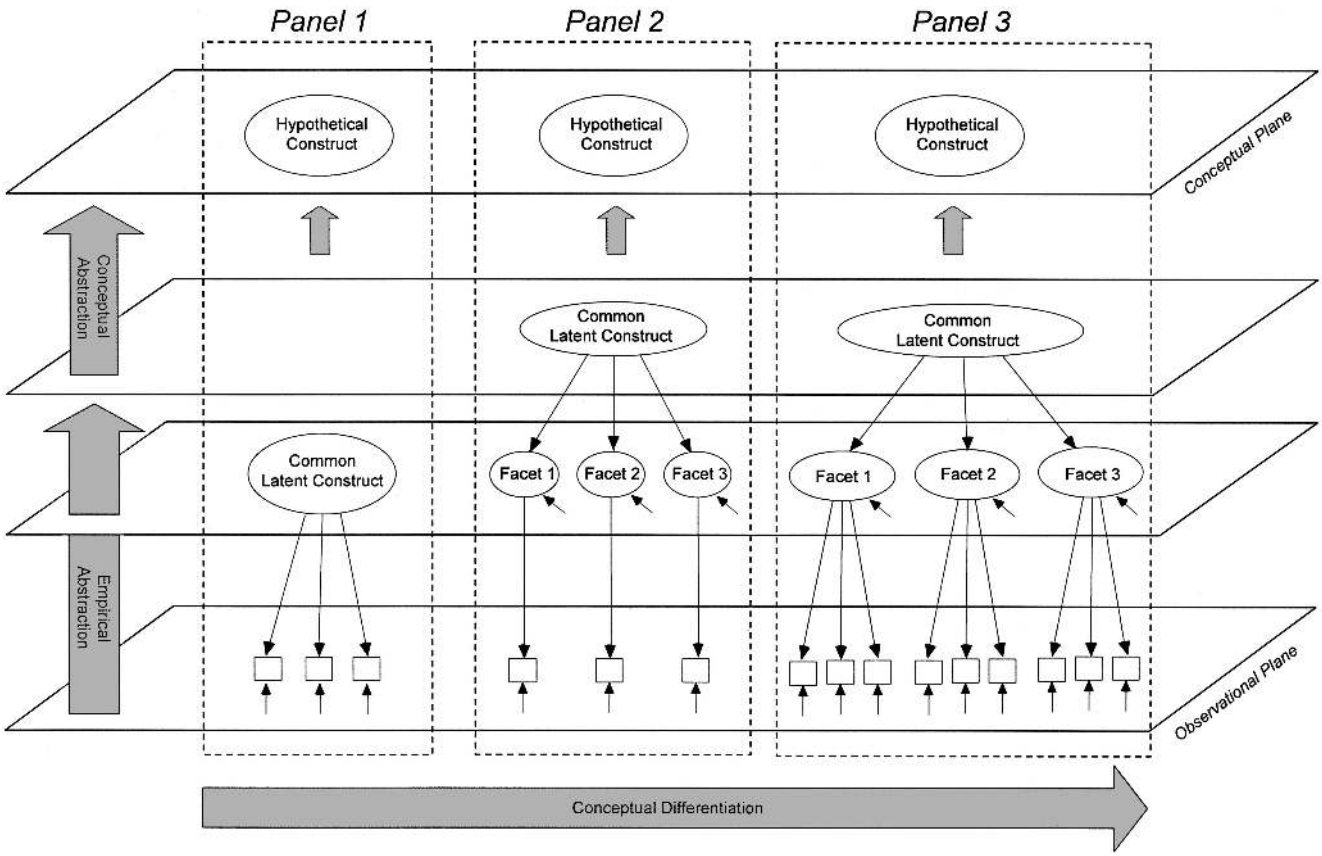

Figure 2. Reflective-indicator measurement models.

domain of this construct as consisting of positive affect toward the supervisor, the desire to interact with him or her, and positive evaluations of him or her. If this construct is not the primary focus of a study, a researcher might measure it in fairly general terms with three items using the measurement model shown in Panel 1 of Figure 2 (e.g., "I really like my supervisor," "I enjoy working on projects with my supervisor," and "I think my supervisor is a nice person"). However, if this construct is the primary focus of the study, it is likely that the researcher will draw sharper conceptual distinctions between the three facets of liking (i.e., affect toward, interaction with, and evaluation of the supervisor) and spend considerably more time and effort on the development and refinement of the measures. In this instance, the researcher may develop multiple items to measure each of the key facets of the construct and use a measurement model like the one shown in Panel 3 of Figure 2. For example, additional measures of the affect subdimension might include items such as "I am very fond of my supervisor" and "I feel positively toward my supervisor"; additional measures of the interaction subdimension might include items like "I frequently talk to my supervisor during breaks and my lunch time" and "I enjoy interacting with my supervisor outside of work"; and additional measures of the evaluation subdimension might include "My supervisor is one of the most pleasant people I know" and "I believe my supervisor is an honest person." The advantage of this measurement model is that it allows one to separate item-level measurement error from measurement error at the level of the subdimension.

The model shown in Panel 2 is a transitional model that is empirically equivalent to the model in Panel 1 (when the item-level measurement error terms are fixed at 0) but represents a conceptual elaboration of it because Panel 2 draws more of a distinction between the facets. The difference between Panels 1 and 2 is that in Panel 2, the researcher has decided that the conceptual distinctions between the facets are important to recognize. This measurement model has also been used as a simplification of the model in Panel 3 when the indicators of each facet are scale scores created by averaging a set of items measuring the facet. In this case, it is conceptually equivalent to the model in Panel 3 but empirically different.

Figure 3 contains a parallel set of interrelated models that have formative indicators of the composite latent construct. Panel 1 shows a composite latent construct with three formative indicators. Panel 3 shows a second-order composite latent construct with three first-order latent constructs as formative indicators, and each of these first-order latent constructs has multiple reflective indicators. Panel 2 is a transitional model that is empirically equivalent to the model in Panel 1, when the item-level measurement error terms are fixed at 0, and conceptually equivalent to the model in Panel 3 because both emphasize the importance of the conceptual distinctions between the facets. Once again, the choice of measurement model would depend on the generality or specificity of one's theoretical interest.

Overall job performance may be a good example of a construct for which these measurement models would be appropriate. Many

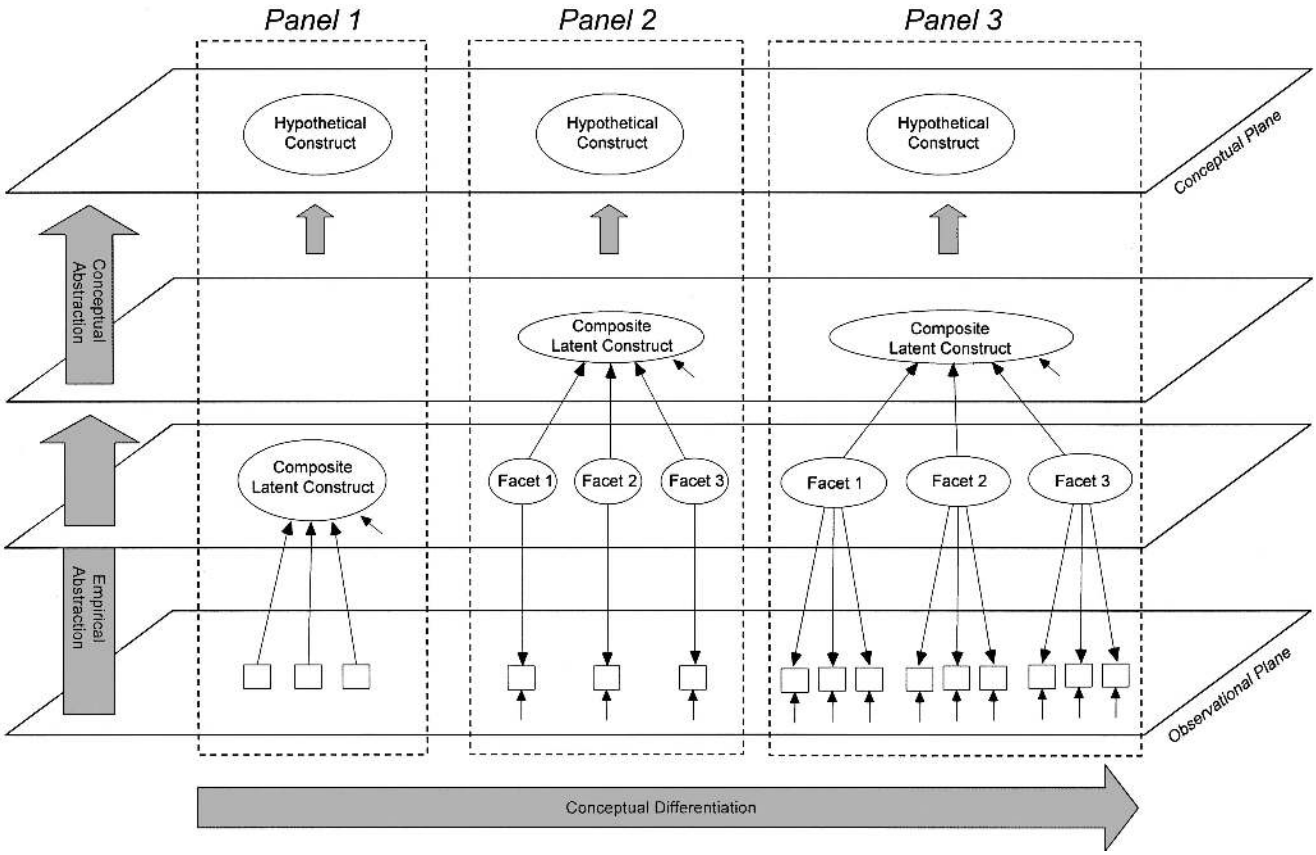

Figure 3. Formative- and mixed-indicator measurement models.

people are now recognizing that job performance is a composite latent construct that has both in-role and extrarole facets. For contextual performance researchers (cf. Borman & Motowidlo, 1993), the facets include task performance, job dedication, and interpersonal facilitation, whereas for organizational citizenship behavior researchers (MacKenzie, Podsakoff, & Ahearne, 1998), the facets include in-role performance, helping behavior, sportsmanship, and civic virtue. However, regardless of the specific approach, all of these researchers agree that overall job performance is a multidimensional construct that comprises several distinct subdimensions. Indeed, we would argue that (a) these subdimensions are all defining characteristics of job performance, because one cannot even think about evaluating job performance without reference to one or more of them; (b) changes in a person's performance in these areas produce changes in his or her job performance; (c) these subdimensions are likely to have different antecedents; and (d) each of these subdimensions captures a unique aspect of the job performance construct domain not captured by the others.

Having said this, it is obvious that there are many different ways in which these subdimensions could be measured—ways that are all equally acceptable from a conceptual point of view. Indeed, Organ (1988) has argued that what we have referred to as helping behavior may be measured differently in different organizational contexts, and Motowidlo (2000) has taken this one step further and argued that this subdimension of contextual performance should be measured differently in different organizational contexts. But both authors view these alternative measures as being equally appropriate reflections of the underlying construct. An employee's scores on the measures would be reflections of the extent to which he or she engaged in helping behavior. Thus, the relations between the measures and the first-order subdimensions of job performance look like those depicted in Panel 3 of Figure 3.

If job performance is the focus of the study, researchers will probably want to use the measurement model shown in Panel 3, because it faithfully represents all of the conceptual distinctions that the researcher believes are important, and it provides the most powerful means of testing and evaluating the construct. In this model, the item-level error terms capture the invalidity and unreliability of the individual measures. This might be caused by contaminating constructs or random factors. However, the error term associated with the composite latent construct of job performance captures the invalidity of the set of subdimensions as measures of the second-order construct. This invalidity may be due to several factors, including the imperfect validity of the individual components or their invalidity as a group due to the failure to include all of the facets that are important aspects of job performance in the measure.

If job performance is less central to the research and/or is part of a complex system of relationships being investigated, researchers might choose to use either the model shown in Panel 1 or the model in Panel 2 with a single measure of each subdimension of

performance. And researchers might use the model in Panel 2 with scale scores as single measures of each subdimension when they wish to have some of the benefits of the complete second-order factor structure (see Panel 3) without drastically increasing the number of measures in their model. However, because these scale scores do not fully represent the subdimensions, there is a trade-off being made. Finally, it is important to recognize that none of the models shown in Figure 3 are identified as they are depicted. Ways of achieving identification in constructs with formative indicators are discussed in the Practical Guidelines for Developing and Evaluating Constructs With Formative Indicators section.

Commonly Misspecified Constructs in Organizational and Behavioral Research

We believe that much of what is said above about job performance also applies to other important constructs in the literature. This is consistent with the view of Law and Wong (1999), who have noted that "the ubiquity of the composite view . . . is evidenced in other constructs such as role conflict and role ambiguity, organizational commitment, occupational health, mental health, and dysfunctional thought processes" (p. 149). Our own reading of the literature suggests that there are many more constructs for which a formative model is appropriate.

For example, another construct that probably should be modeled as a composite latent construct is transformational leadership. This construct is often conceptualized as being a function of charisma, idealized influence, inspirational leadership, intellectual stimulation, and individualized consideration (cf. Bass, 1985, 1998). In our view, these forms of leader behavior are conceptually distinct, likely to have different antecedents and/or consequences, and are not interchangeable. Indeed, it is not difficult to imagine a leader who is able to demonstrate consideration to followers (e.g., exhibit individualized consideration) but is not able to get them to question the appropriateness of critical assumptions they have about their work (e.g., exhibit intellectual stimulation) or able to display a sense of power and confidence (e.g., exhibit idealized influence). Thus, even though this construct has consistently been modeled in the literature as having reflective indicators (cf. Bycio, Hackett, & Allen, 1995; Geyer & Steyrer, 1998; Tracey & Hinkin, 1998), Bass's (1985) transformational leadership construct should be modeled as having formative indicators, probably as shown in Panel 3 of Figure 3. The same can be said for Podsakoff, MacKenzie, Moorman, and Fetter's (1990) slightly different conceptualization of transformational leadership.

Another example of a construct that should be modeled as having formative indicators is procedural justice. According to Colquitt (2001), procedural justice consists of the perceived fairness of the procedures used to arrive at a person's job outcomes, including whether (a) the procedures are developed with the employee's input, are applied consistently, are free of bias, and are based on accurate information; (b) employees have influence over the outcome; and/or (c) employees have the ability to appeal the outcome. Clearly, perceptions of the overall fairness of the procedures are the result of these things rather than the cause of them. Moreover, the items are not interchangeable, and because all of the characteristics mentioned in them are necessary for the procedures to be perceived as fair, eliminating one or more of the items would alter the conceptual domain of the construct and undermine its validity. Finally, some of the things that one might do to make sure that the procedures are free of bias and consistently applied would be quite different from what one might do to ensure that the procedures are developed with the employee's input. Therefore, contrary to how it has been modeled in the literature (cf. Colquitt, 2001; Masterson, 2001; Moorman, 1991), procedural justice should probably be modeled as having formative indicators.

Of course, we do not mean to imply that every construct, or even most constructs, should be modeled as having formative indicators. Nor are we suggesting that the authors cited in the above examples should have been aware of the distinction between formative and reflective indicators at the time their research was conducted. Instead, our point is that the specification of the measurement model is a critical decision that needs to be made on the basis of conceptual criteria like the ones that we have discussed.

The Severity of the Effects of Misspecification

The preceding examples demonstrate that measurement model misspecification is fairly common among published research studies. This is an important problem, because empirical research (Law & Wong, 1999) has demonstrated that measurement model misspecification can bias structural parameter estimates. In Law and Wong's study, the parameter estimate for the relationship between job perception and job satisfaction was inflated by 132% (.995 vs. .429) when job perception was incorrectly modeled as having reflective indicators compared with when it was correctly modeled as having formative indicators. However, both estimates were statistically significant, so no error of inference was made. Law and Wong also found that measurement model misspecification in one construct can sometimes influence relationships in a model that do not involve this construct and result in errors of inference.

Although Law and Wong (1999) provided an important demonstration of the potential effects of measurement model misspecification, there are three factors that limit the extent to which the findings of their study can be generalized. First, because their results were only based on data from a single sample, it is difficult to know whether the results might be sample specific. Second, because Law and Wong's sample was relatively small ($N = 204$), and error rates are known to be sensitive to sample size, it is not clear how generalizable their findings about the effects of measurement model misspecification on error rates might be. Finally, the generalizability of their results may also be limited by the fact that the error term for the job perception construct was not identified when job perception was specified as having formative indicators. This is because the job perception construct did not have two paths emanating from it that led to independent constructs. It had two paths leading from it, but the two constructs were causally related. This causes the error term for the composite latent construct (job perception) and the structural error terms for the constructs it influences to be indeterminate (i.e., unique parameter estimates are not obtainable). This is an identification problem that was explicitly noted by MacCallum and Browne (1993).

Simulation Objectives

In view of the above limitations, there is still not a very clear picture of how severe the effects of measurement model misspeci-

fication might be. Therefore, we conducted a Monte Carlo simulation designed to investigate three unresolved issues. First, to what extent is the estimate of a structural relationship between two constructs (e.g., γ) biased by measurement model misspecification? We expected that the position of the misspecified construct in a model (e.g., exogenous vs. endogenous) would influence the degree of bias observed. This is because treating the formative indicators of a construct as if they were reflective indicators reduces the variance of the latent construct, because this defines the construct in terms of the common variance of the indicators rather than in terms of their total variance (cf. Law & Wong, 1999, p. 145). When the misspecified construct is in the exogenous position, the variance of the exogenous construct will go down, thus resulting in an upward bias in the estimates of the impact of this construct on other constructs because the estimate captures the effect of a one-unit change in the exogenous construct on the endogenous construct. Conversely, when the misspecified construct is in the endogenous position, the variance of the endogenous construct will be reduced, thereby producing a downward bias in the structural estimate of the relationship between this construct and an exogenous construct. However, when both the endogenous and exogenous constructs are misspecified, this reasoning would lead to the prediction that effects would tend to partially cancel each other out, thus producing less of a bias. The magnitude of the item intercorrelations was also expected to influence the extent of the bias in the structural estimates, because the greater the magnitude of the item intercorrelations, the smaller the change in the variance of a construct produced by measurement model misspecification. Therefore, the greater the magnitude of the item intercorrelations, the smaller the bias in the structural estimates produced by measurement model misspecification. In this sense, the magnitude of the interitem correlations captures the degree of measurement model misspecification, with high item intercorrelations indicating a less severe misspecification than lower intercorrelations.

The second unresolved question that the simulation was designed to address is this: To what extent will measurement model misspecification lead to errors of inference in hypothesis testing? The statistical inferences about relationships between constructs are based on the critical ratio of the magnitude of the unstandardized structural estimate and the standard error of that estimate. Therefore, things that bias either of these two values have the potential to lead to errors. It is widely recognized that sample size is inversely related to the magnitude of the standard error of the structural estimate (Cohen, 1988). In addition, we have already argued that the position of a misspecified construct in a model (e.g., exogenous or endogenous) and the strength of the item intercorrelations will bias the magnitude of the structural parameter estimates. Consequently, these factors should also influence the error rate by influencing either the numerator or the denominator of the critical ratio. More specifically, we expected that (a) misspecification of the exogenous construct would inflate the structural parameter estimate, thus increasing Type I and decreasing Type II error rates; (b) misspecification of the endogenous construct would deflate the structural parameter estimate, thus decreasing Type I and increasing Type II error rates; (c) as the magnitude of the item intercorrelations increased, the bias in the structural estimates produced by either type of measurement model misspecification would decrease, and the effects of these factors on the Type I and Type II error rates should decrease; and (d) increasing the sample size would decrease the standard error of the structural parameter estimate and increase Type I and decrease Type II error rates. In addition, although we were not necessarily predicting that these factors would interact to influence the error rate observed, this possibility was examined in the simulation.

The third question addressed by the simulation is this: To what extent are the most commonly used goodness-of-fit indices capable of detecting measurement model misspecification? This research used the goodness-of-fit index (GFI), comparative fit index (CFI), standardized root-mean-square residual (SRMR), and root-mean-square error of approximation (RMSEA) as the indices of model fit. The GFI was selected because it is one of the most widely reported indices in the literature; the SRMR was selected because Hu and Bentler (1998) found it to be the most sensitive goodness-of-fit index for detecting misspecified relationships between latent constructs; and the CFI and RMSEA were selected because Hu and Bentler (1998) found that they were the most sensitive goodness-of-fit indices at detecting measurement model misspecification. Because high CFI and GFI values and low SRMR and RMSEA values are associated with better fitting models, we expected the CFI and GFI to be negatively related to the degree of model misspecification and the SRMR and RMSEA to be positively related to the degree of model misspecification. This means that we should have observed main effects of the endogenous and exogenous construct misspecification manipulations and perhaps an interaction between the two on these goodness-of-fit indices. Similarly, because the degree of measurement model misspecification is greater when the item intercorrelations are low than when they are high, we expected that our manipulation of the magnitude of the item intercorrelations would have a main effect on the goodness-of-fit indices. More specifically, as intercorrelations among the items increased, the CFI and GFI should have increased, and the SRMR and RMSEA should have decreased. Finally, on the basis of previous simulation research (cf. Anderson & Gerbing, 1984; Bearden, Sharma, & Teel, 1982; Bollen, 1990; Hu & Bentler, 1998, 1999), we expected that sample size would be positively related to the GFI, negatively related to the SRMR and RMSEA fit indices, and generally not related to the CFI. However, it is important to note that with the exception of Hu and Bentler (1998), the effects of model misspecification on this pattern of results have not been examined.

Simulation Design

Therefore, we conducted a Monte Carlo simulation to investigate these issues. More specifically, the simulation examined the empirical consequences of inappropriately applying a reflective-indicator measurement model to a construct that should have been modeled with formative indicators. Figure 4 summarizes the models tested in the Monte Carlo simulation conditions. The simulation conditions varied on the basis of whether (a) the measurement model of the focal construct was correctly specified as having formative indicators (as indicated in Figure 4A) or incorrectly specified as having reflective indicators (as indicated in Figures 4B, 4C, and 4D); (b) the size of the sample used in the simulation was small (125), medium (500), or large (875); (c) the item intercorrelations of the focal construct were relatively weak (.20), moderate (.50), or strong (.80); and (d) the relationship between

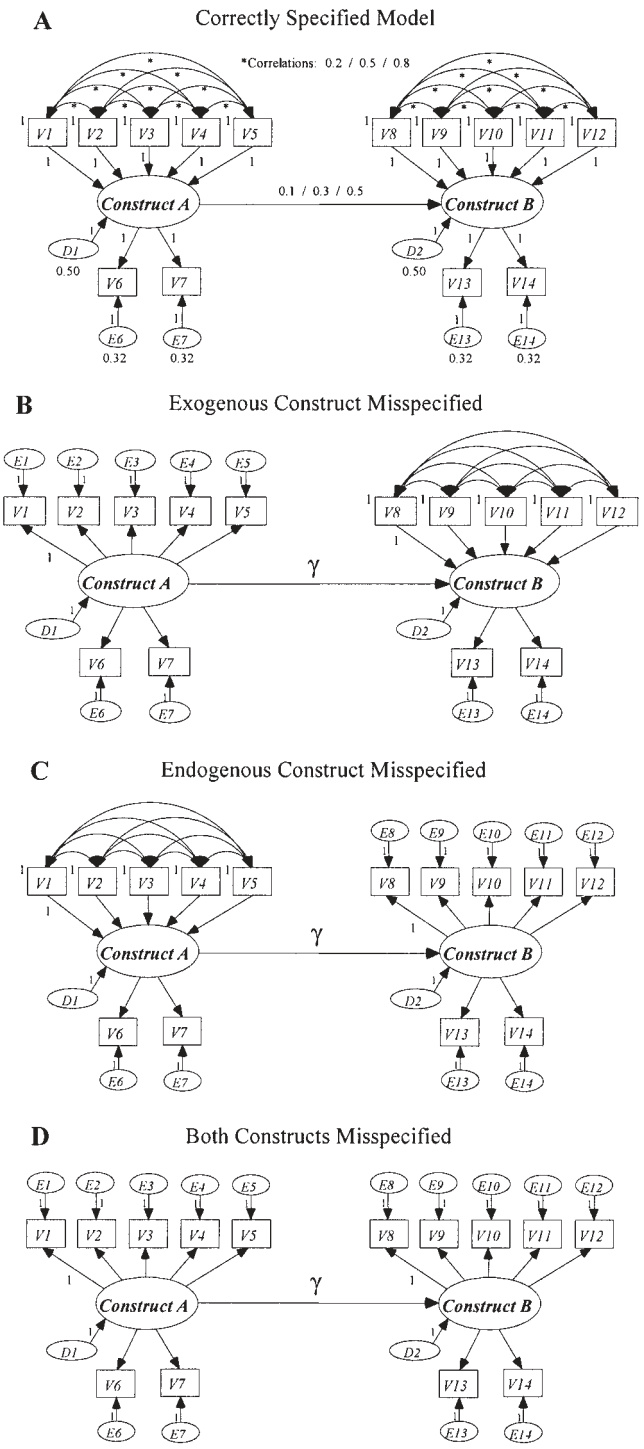

Figure 4. Summary of simulation models.

the exogenous and endogenous constructs was weak (e.g., .10), moderate (e.g., .30), or strong (e.g., .50). It is important to note two things about these models. First, the correctly specified model (see Figure 4A) includes two reflective indicators in addition to its five formative indicators. As we elaborate on in our discussion of Figure 8 (presented later), these reflective indicators were included to ensure that the otherwise formatively measured construct would be identified (cf. MacCallum & Browne, 1993). Second, we did not include cells in the simulation design where there was no relationship between the exogenous and endogenous constructs because we wanted to focus on how measurement model misspecification influences estimates of structural relationships of known strength and statistical tests of those relationships, and we wanted to do this under conditions in which the power was great enough that any errors of inference (i.e., Type II errors) would be due primarily to the manipulated factors rather than chance.

Manipulating whether the exogenous or endogenous constructs were misspecified allowed us to examine the consequences of measurement model misspecification on structural paths emanating from and going into the misspecified constructs. This is important because both types of misspecification are commonly found in the literature, and their effects are expected to be different. The manipulation of the magnitude of the correlations among the indicators of the construct with the formative indicators allowed us to test the significance of the effects of misspecification across a variety of situations, including some in which the item intercorrelations were high enough (e.g., .50 and .80) that it might appear that a reflective-indicator model is appropriate. Indeed, a five-item scale with item intercorrelations of .50 would have a Cronbach's alpha of .83, and with intercorrelations of .80, it would have a Cronbach's alpha of .95. Manipulating sample size in the simulation permitted us to examine the effects of the size of the sample on the likelihood of Type II errors and the goodness-of-fit indices. The mean sample size in our simulation (500) was selected to correspond to the mean sample size in articles published in several organizational and management journals from 1995–1997 (cf. Scandura & Williams, 2000); the small sample size (125) was selected to roughly correspond to the minimum needed to safely meet the sample-size requirements for the maximum-likelihood estimation method; and the large sample size (875) was selected to be an equal distance from the mean sample size and to roughly correspond to a typical "large" sample. Finally, the manipulation of the strengths of the paths emanating from the construct with formative indicators allowed us to evaluate the consequences of measurement model misspecification across the range of relationship strengths typically reported in psychological research (cf. Cohen, 1992).

After the population values were set, nine population covariance matrices were calculated for the true model (see Figure 4A)—one for each of the unique combinations of interitem correlation strength (.20, .50, .80) and relationship strength (.10, .30, .50). Following this, a Monte Carlo simulation was conducted using EQS 5.7b (Multivariate Software, Inc., Encino, CA). The simulation generated raw data sets of varying sample size (125, 500, 875) from each of the population covariance matrices, assuming a normal distribution. The data were sampled from a normal distribution because the maximum-likelihood technique used to estimate the model assumes multivariate normality and because most other researchers have used normally distributed data when conducting statistical equation modeling simulation studies. A total of 500 data sets were generated for each condition. These data sets were then fit to the models shown in Figure 4, generating parameter estimates and fit statistics for each replication. We were particularly interested in the unstandardized structural parameter

estimate (γ), its standard error, and the overall indices of model fit (CFI, GFI, RMSEA, and SRMR).

Simulation Results

Before turning to the results of the simulation study, it is important to note that with 500 cases per cell and 108 cells, the power to detect statistically significant effects of the manipulations was quite large. Indeed, according to Cohen (1988), with our sample size and using an alpha level of $p < .05$, there was virtually a 100% chance of detecting a small, medium, or large effect. Consequently, most of the manipulated effects and their interactions were statistically significant ($p < .05$). However, this does not mean that they were all equally important in accounting for variation in the criterion measures. Therefore, we calculated partial η^2 estimates to identify the most important effects, and only those effects that accounted for approximately 5% of the variance in the criterion measure are discussed.

Table 1 reports the means for the criterion measures for the simulation treatment conditions, and Table 2 reports the analysis of variance results. The first four columns of Table 1 indicate the treatment condition. More specifically, the first two columns indicate whether the endogenous and/or exogenous constructs were misspecified, the third column indicates the interitem correlation values, and the fourth column indicates the sample size of each of the treatment conditions. Column 1 in Table 2 reports the main and interactive effects of the manipulations of the misspecification of the endogenous construct, exogenous construct, item intercorrelations, and sample size. Although it was important for generalizability purposes to also manipulate the magnitude of the effect of the exogenous construct on the endogenous construct (γ) across a range of effect sizes, the results shown in Tables 1 and 2 are collapsed across these conditions for several reasons. First, the effect of this manipulation on the magnitude of the γ coefficient is theoretically uninteresting, because this factor is a direct manipu-

Table 1
Criterion-Measure Means for All Treatment Conditions of the Monte Carlo Simulation

Treatment condition				Measure					
Endogenous misspecified	Exogenous misspecified	Interitem correlation	Sample size	γ	*SE* of γ	CFI	GFI	RMSEA	SRMR
No	No	.20	125	0.299	0.024	.996	.937	.021	.060
No	No	.20	500	0.301	0.012	.999	.984	.008	.030
No	No	.20	875	0.300	0.009	1.000	.991	.006	.023
No	No	.50	125	0.301	0.019	.997	.937	.022	.058
No	No	.50	500	0.300	0.010	.999	.984	.009	.030
No	No	.50	875	0.300	0.007	1.000	.991	.006	.022
No	No	.80	125	0.300	0.016	.998	.937	.021	.055
No	No	.80	500	0.300	0.008	1.000	.984	.008	.028
No	No	.80	875	0.300	0.006	1.000	.991	.006	.021
No	Yes	.20	125	1.717	0.270	.913	.880	.123	.075
No	Yes	.20	500	1.613	0.126	.916	.927	.121	.052
No	Yes	.20	875	1.608	0.095	.916	.934	.121	.047
No	Yes	.50	125	1.606	0.165	.960	.891	.098	.064
No	Yes	.50	500	1.564	0.081	.962	.939	.096	.038
No	Yes	.50	875	1.559	0.061	.962	.946	.096	.032
No	Yes	.80	125	1.549	0.109	.993	.917	.045	.055
No	Yes	.80	500	1.540	0.054	.995	.968	.043	.028
No	Yes	.80	875	1.543	0.041	.995	.976	.044	.022
Yes	No	.20	125	0.046	0.017	.897	.876	.134	.090
Yes	No	.20	500	0.047	0.009	.898	.922	.133	.077
Yes	No	.20	875	0.047	0.007	.899	.929	.133	.074
Yes	No	.50	125	0.047	0.017	.927	.860	.131	.105
Yes	No	.50	500	0.048	0.008	.929	.903	.129	.096
Yes	No	.50	875	0.048	0.006	.929	.910	.129	.094
Yes	No	.80	125	0.046	0.017	.932	.767	.146	.122
Yes	No	.80	500	0.046	0.008	.932	.797	.145	.119
Yes	No	.80	875	0.046	0.006	.932	.802	.145	.118
Yes	Yes	.20	125	0.263	0.103	.813	.826	.170	.101
Yes	Yes	.20	500	0.251	0.050	.814	.872	.169	.088
Yes	Yes	.20	875	0.253	0.037	.814	.879	.169	.086
Yes	Yes	.50	125	0.246	0.090	.889	.821	.153	.108
Yes	Yes	.50	500	0.249	0.045	.892	.866	.151	.098
Yes	Yes	.50	875	0.247	0.034	.892	.873	.151	.097
Yes	Yes	.80	125	0.234	0.086	.927	.754	.143	.122
Yes	Yes	.80	500	0.239	0.042	.928	.788	.142	.118
Yes	Yes	.80	875	0.235	0.032	.929	.798	.141	.116

Note. All means have been collapsed across levels of gamma. Endogenous = endogenous construct; Exogenous = exogenous construct; γ = unstandardized structural parameter estimate; CFI = comparative fit index; GFI = goodness-of-fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.

Table 2
Analysis of Variance Results for the Monte Carlo Simulation

Source	*df*[a]	gamma			*SE* of gamma			CFI			GFI			RMSEA			SRMR		
		F	*p*	P η^2	*F*	*p*	P η^2	*F*	*p*	P η^2	*F*	*p*	P η^2	*F*	*p*	P η^2	*F*	*p*	P η^2
Endogenous	1	38,999.90	.00	.21	14,065.20	.00	.05	151,492.00	.00	.32	46,926.10	.00	.26	160,300.00	.00	.35	47,276.90	.00	.29
Exogenous	1	33,926.10	.00	.19	96,695.90	.00	.27	42,100.10	.00	.12	5,582.21	.00	.04	38,540.60	.00	.12	624.56	.00	.01
Interitem correlation value	2	18.93	.00	.00	4712.72	.00	.03	26,712.40	.00	.14	2,922.47	.00	.04	3,099.87	.00	.02	439.30	.00	.01
Sample size	2	4.74	.01	.00	15,776.00	.00	.11	42.64	.00	.00	4,241.00	.00	.06	160.43	.00	.00	2,238.68	.00	.04
Endogenous × Exogenous	1	18,169.80	.00	.11	12,178.80	.00	.04	0.04	.83	.00	59.45	.00	.00	14,287.30	.00	.05	97.72	.00	.00
Endogenous × Correlation Value	2	9.07	.00	.00	3,107.11	.00	.02	2,687.55	.00	.02	6,135.46	.00	.08	1,840.79	.00	.01	2,318.00	.00	.04
Endogenous × Sample Size	2	4.16	.02	.00	1,397.07	.00	.01	0.50	.61	.00	23.24	.00	.00	74.58	.00	.00	622.38	.00	.01
Exogenous × Correlation Value	2	18.83	.00	.00	3,797.47	.00	.03	12,471.20	.00	.07	625.91	.00	.01	5,128.06	.00	.03	273.67	.00	.00
Exogenous × Sample Size	2	5.00	.01	.00	9,340.97	.00	.07	0.28	.76	.00	3.62	.03	.00	73.68	.00	.00	4.36	.01	.00
Correlation Value × Sample Size	4	1.65	.16	.00	531.52	.00	.01	1.13	.34	.00	5.68	.00	.00	0.20	.94	.00	11.81	.00	.00
Endogenous × Exogenous × Correlation Value	2	9.08	.00	.00	2,503.48	.00	.02	0.97	.38	.00	0.20	.82	.00	589.41	.00	.00	15.51	.00	.00
Endogenous × Exogenous × Sample Size	2	3.92	.02	.00	1,244.46	.00	.01	1.25	.29	.00	0.20	.82	.00	75.49	.00	.00	8.81	.00	.00
Endogenous × Correlation Value × Sample Size	4	0.96	.43	.00	336.56	.00	.01	0.56	.69	.00	10.97	.00	.00	0.21	.93	.00	13.78	.00	.00
Exogenous × Correlation Value × Sample Size	4	1.73	.14	.00	441.49	.00	.01	0.36	.84	.00	1.64	.16	.00	0.27	.90	.00	3.15	.01	.00
Endogenous × Exogenous × Correlation Value × Sample Size	4	0.98	.42	.00	278.60	.00	.00	0.12	.98	.00	0.18	.95	.00	0.14	.97	.00	1.14	.34	.00

Note. P = partial; Endogenous = endogenous construct misspecified; Exogenous = exogenous construct misspecified; CFI = comparative fit index; GFI = goodness-of-fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.
[a] Error *df* for all *F*s = 51,244.

lation of the magnitude of this coefficient. Therefore, testing it would simply reveal that our manipulation worked. Second, this factor was not expected to influence the other criterion measures (CFI, GFI, RMSEA, and SRMR). Indeed, the only consistent effect of the γ manipulation was that the overall goodness-of-fit indices decreased as the γ value increased, but only when the endogenous construct was misspecified. Third, the addition of this factor would have tripled the size of Table 1 and obscured the more important relationships in the data.

Effects of the manipulations on the structural parameter estimate and Type II error rate. As indicated in Table 2, the magnitude of the structural parameter estimate (γ) was primarily determined by the endogenous and exogenous construct manipulations and their interactions. These manipulations accounted for 21%, 19%, and 11%, respectively, of the variance of this criterion variable. These same manipulations and their interactions accounted for 5%, 27%, and 4%, respectively, of the variance in the standard error of the structural parameter estimate. These effects are depicted in Figure 5. Figure 5A shows that when the exogenous construct alone was misspecified, the structural parameter estimate (1.59) was 429% higher than its true value of 0.30. However, when the endogenous construct alone was misspecified, the structural parameter estimate (0.048) was 84% lower than its true value of 0.30. Finally, when both the endogenous and exogenous constructs were misspecified, the structural parameter estimate (0.246) was 18% lower than its true value of 0.30. Figure 5B shows that misspecification of the exogenous construct greatly inflates the standard error of the estimate, misspecification of the endogenous construct slightly deflates the standard error of the estimate, and misspecification of both constructs has the net effect of inflating the standard error. Thus, measurement model misspecification was found to either inflate or deflate the structural parameter estimate and the standard error of the estimate, depending on whether the exogenous or endogenous construct was misspecified.

We investigated the extent to which misspecification influenced inferences about the statistical significance of this relationship (i.e., the Type II error rate) using logit regression analysis by calculating the critical ratio of the unstandardized estimate to its standard error and examining whether the ratio was greater or less than the critical value of 1.96 (for $\alpha < .05$). When the exogenous construct alone was misspecified, the average Type II error rate was found to be 0. When the endogenous construct alone was misspecified, the average Type II error rate was 19%. Finally, when both constructs were misspecified, the average Type II error rate was 18%. This means that the statistical test falsely indicated that there was no relationship between the two constructs when there really was one 19% of the time when the endogenous construct was misspecified, even though the power was sufficient to detect even the weakest relationship virtually 100% of the time. Thus, consistent with our expectations, (a) misspecification of the exogenous construct inflated the structural parameter estimate but did not reduce the Type II error rate (because it was already at 0), and (b) misspecification of the endogenous construct deflated the structural parameter estimate and, consequently, inflated the Type II error rate. Although we did not have specific hypotheses regarding the interactive effects, misspecifying both the exogenous and endogenous constructs had almost exactly the same inflationary effect on the Type II error rate as misspecifying the endogenous construct alone (18% vs. 19%, respectively). However, as shown in Figure 5, when both constructs were misspecified, the increase in the Type II error rate was due much more to an inflation of the standard error of the estimate than to a deflation of the estimate itself, whereas when only the endogenous construct was misspecified, the increase in the Type II error rate was due more to a deflation of the estimate than to an inflation of its standard error.

In addition to the effects discussed above, the standard error of the estimate was also influenced by sample size (partial $\eta^2 = 11\%$), as expected. As indicated in Table 1, the standard error of the estimate decreased as sample size increased. Moreover, this effect was found to be stronger when the exogenous construct was misspecified than when it was correctly specified (partial $\eta^2 = 7\%$).

Finally, as expected, the bias in the structural parameter estimate (γ) produced by the misspecification of the exogenous or endogenous constructs decreased as the item intercorrelations increased. The same was also true for the bias in the standard error of the estimate that was produced by the misspecification of the endogenous or exogenous constructs. However, contrary to expectations, the Type II error rates did not follow this pattern, because the rates of decline of the structural parameter estimate and its standard error were different. The effects of misspecification of the exogenous construct on the Type II error rate decreased only trivially as the item intercorrelations increased, whereas the effects of misspecification of the endogenous construct on the Type II error rate actually increased slightly as the item intercorrelations increased. Thus, the findings provide only partial support for the expected effects; however, it is important to note that none of these interaction effects accounted for a very substantial percentage of the variance (see Table 2).

Effects of the manipulations on goodness of fit. Generally speaking, we expected to observe endogenous, exogenous, and item intercorrelation main effects (and perhaps their interactions) on the goodness-of-fit indices. We also generally expected sample size to influence them, although less so for the CFI. Consistent with these expectations, Tables 1 and 2 indicate that when the endogenous construct was misspecified, the CFI decreased (partial $\eta^2 = 32\%$); when the exogenous construct was misspecified, the CFI decreased (partial $\eta^2 = 12\%$); and as the interitem correlation values decreased, the CFI decreased (partial $\eta^2 = 14\%$). In addition, Figure 6A shows that there was also an interaction between the manipulation of the exogenous construct specification and the item intercorrelation that accounted for 7% of the variance in this criterion variable. As indicated in this figure, misspecification of the measurement model for the exogenous construct decreased the CFI much more when the item intercorrelations were low than when they were high. This makes sense, because both of these manipulations influenced the severity of the model misspecification.

Misspecifying the endogenous construct decreased the GFI (partial $\eta^2 = 26\%$), and the GFI also decreased as the sample size decreased (partial $\eta^2 = 6\%$). In addition, as shown in Figure 6B, there was an interaction between the manipulation of the specification of the endogenous construct and the interitem correlation value (partial $\eta^2 = 8\%$). The GFI was positively related to the interitem correlation value when the endogenous construct was not misspecified, but it was negatively related to the interitem correlation value when the endogenous construct was misspecified. With respect to the RMSEA, the results indicated that this criterion

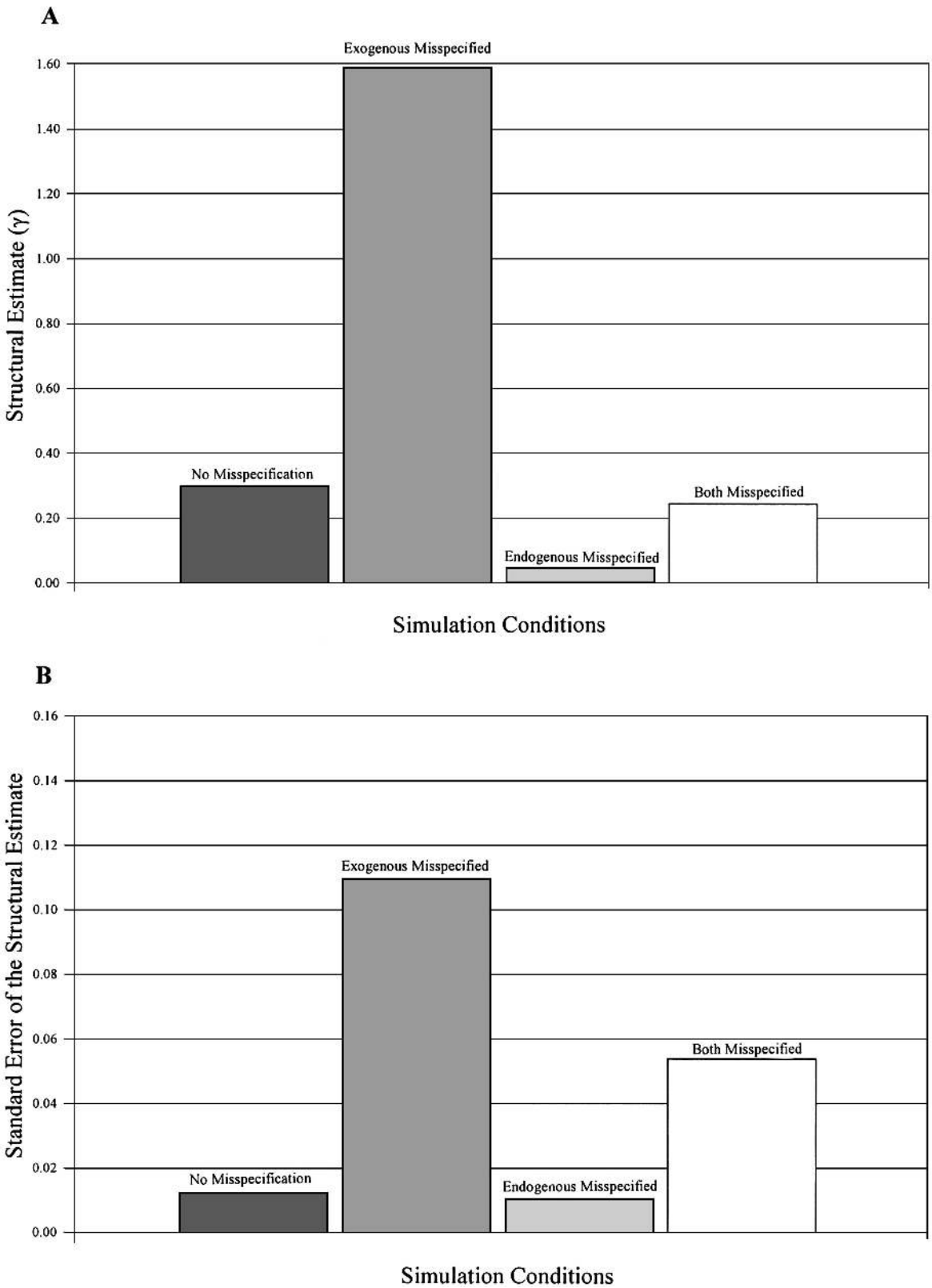

Figure 5. Illustrations of the effects of exogenous and endogenous construct misspecification on the structural parameter estimate (A) and the standard error of the structural parameter estimate (B).

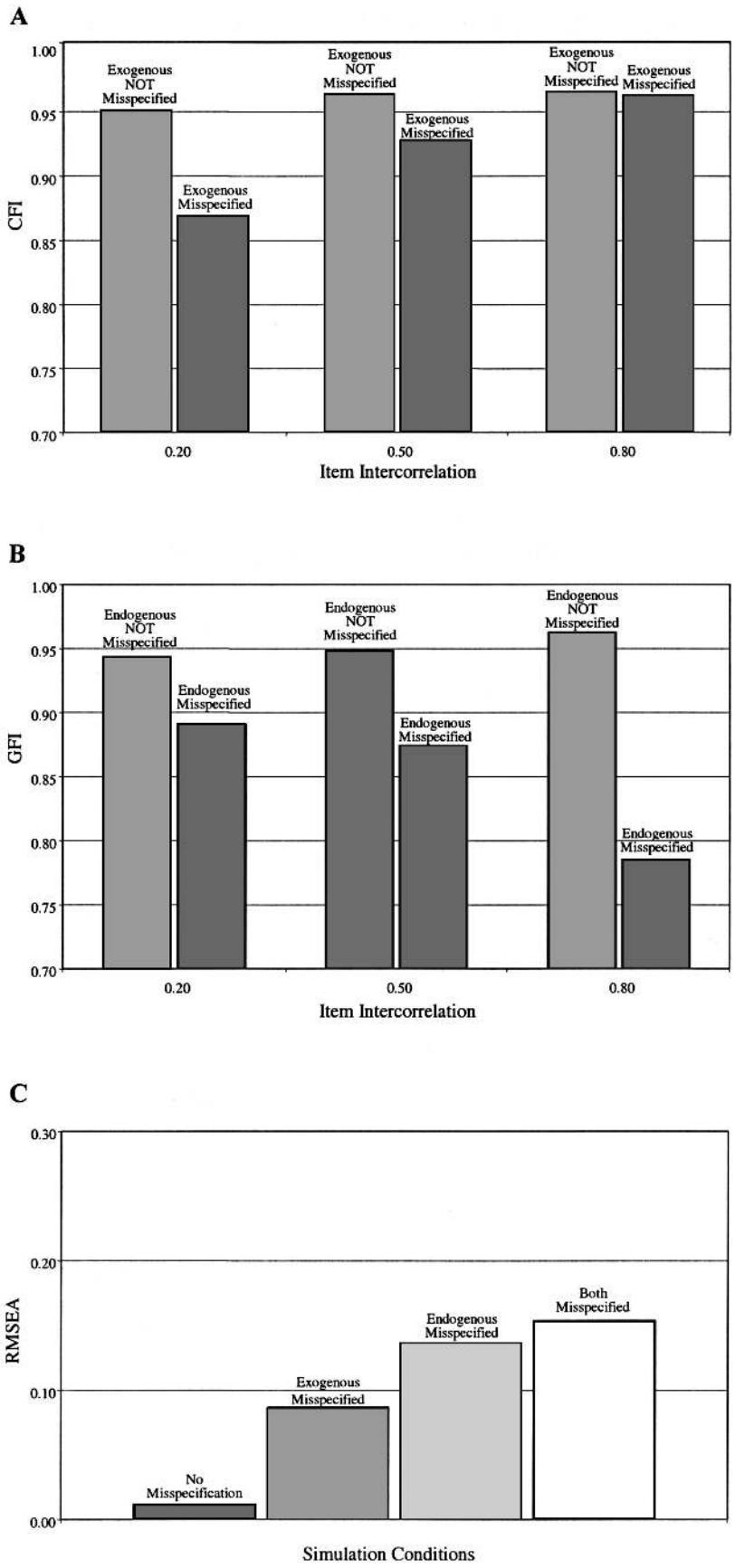

Figure 6. Illustrations of interaction effects on (A) the comparative fit index (CFI), (B) the goodness-of-fit index (GFI), and (C) the root-mean-square error of approximation (RMSEA).

variable increased when the exogenous construct was misspecified (partial $\eta^2 = 12\%$) and when the endogenous construct was misspecified (partial $\eta^2 = 35\%$). In addition, there was also a fairly substantial interaction effect between these two factors (partial $\eta^2 = 5\%$), depicted in Figure 6C. As shown in the figure, the RMSEA increased as the degree of measurement model misspecification increased, but not as much as would be expected if the effects of the endogenous and exogenous misspecification effects were simply additive. Finally, the results indicated that misspecifying the endogenous construct increased the SRMR (partial $\eta^2 = 29\%$).

In our discussion so far, we have focused on the effects that the manipulations had on the means of the goodness-of-fit indices. However, one issue that has not been addressed is the extent to which the goodness-of-fit indices are capable of detecting measurement model misspecification. This can be evaluated by calculating the percentages of time that the fit indices failed to detect measurement model misspecification using traditionally accepted cutoff criteria for each of the goodness-of-fit indices (cf. Hu & Bentler, 1998, 1999): .95 for the CFI, .90 for the GFI, .05 for the RMSEA, and .08 for the SRMR. Table 3 shows the percentages of time that each of the four goodness-of-fit indices failed to detect the misspecification of the measurement model. As shown in the table, the best goodness-of-fit-index at detecting measurement model misspecification was the RMSEA, followed by the CFI. These findings are consistent with Hu and Bentler (1998). The RMSEA successfully detected the measurement model misspecification, except when the exogenous variable was misspecified and the item intercorrelations were high (e.g., .80.) The CFI successfully detected measurement model misspecification in all cases in which the item intercorrelations were low (.20), and it was better at detecting misspecification in the endogenous construct than in

Table 3
Error Rate Percentages of Incorrect Inferences About Model Goodness of Fit

Treatment condition				Measure			
Endogenous misspecified	Exogenous misspecified	Interitem correlation	Sample size	CFI	GFI	RMSEA	SRMR
No	No	.20	125	0%	0%	10%	7%
No	No	.20	500	0%	0%	0%	0%
No	No	.20	875	0%	0%	0%	0%
No	No	.50	125	0%	0%	10%	12%
No	No	.50	500	0%	0%	0%	0%
No	No	.50	875	0%	0%	0%	0%
No	No	.80	125	0%	0%	10%	16%
No	No	.80	500	0%	0%	0%	0%
No	No	.80	875	0%	0%	0%	0%
No	Yes	.20	125	0%	3%	0%	71%
No	Yes	.20	500	0%	100%	0%	100%
No	Yes	.20	875	0%	100%	0%	100%
No	Yes	.50	125	89%	21%	0%	84%
No	Yes	.50	500	100%	100%	0%	100%
No	Yes	.50	875	100%	100%	0%	100%
No	Yes	.80	125	100%	93%	56%	84%
No	Yes	.80	500	100%	100%	85%	100%
No	Yes	.80	875	100%	100%	94%	100%
Yes	No	.20	125	0%	1%	0%	32%
Yes	No	.20	500	0%	100%	0%	61%
Yes	No	.20	875	0%	100%	0%	66%
Yes	No	.50	125	26%	6%	0%	33%
Yes	No	.50	500	33%	67%	0%	33%
Yes	No	.50	875	33%	67%	0%	33%
Yes	No	.80	125	35%	22%	3%	33%
Yes	No	.80	500	33%	33%	0%	33%
Yes	No	.80	875	33%	33%	0%	33%
Yes	Yes	.20	125	0%	0%	0%	8%
Yes	Yes	.20	500	0%	0%	0%	34%
Yes	Yes	.20	875	0%	0%	0%	33%
Yes	Yes	.50	125	0%	0%	0%	32%
Yes	Yes	.50	500	0%	8%	0%	34%
Yes	Yes	.50	875	0%	33%	0%	34%
Yes	Yes	.80	125	34%	4%	0%	34%
Yes	Yes	.80	500	34%	34%	0%	34%
Yes	Yes	.80	875	35%	35%	0%	35%

Note. Errors of inference in the first nine rows refer to the percentages of time that the correctly specified model was falsely rejected as being inconsistent with the data. In the remaining rows, errors of inference refer to the percentages of time that a misspecified model was falsely judged to fit the data. Endogenous = endogenous construct; Exogenous = exogenous construct; CFI = comparative fit index; GFI = goodness-of-fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.

the exogenous construct. The goodness-of-fit index that was the least sensitive at detecting measurement model misspecification was the SRMR. Generally speaking, it performed especially poorly when the exogenous variable was misspecified, and it was incorrect about a third of the time when the endogenous variable was misspecified. Overall, with the exception of the RMSEA, the goodness-of-fit indices tested will fail to detect measurement model misspecification a substantial proportion of the time.

Practical Guidelines for Developing and Evaluating Constructs With Formative Indicators

In view of the fact that some of the most widely researched constructs in the literature are misspecified and that measurement model misspecification can have potentially serious effects, it appears that the field would benefit from some practical guidelines for developing, modeling, and evaluating constructs with formative indicators. Figure 7 provides an overview of the similarities and differences in the stages of the scale development and validation process for constructs with reflective indicators versus constructs with formative indicators. Most of the stages in this process have been described elsewhere (cf. Churchill, 1979; Nunnally & Bernstein, 1994; Schwab, 1980) and are elaborated on here only when the procedures that one should follow for developing and evaluating formative-indicator models differ from those that should be followed for reflective-indicator models.

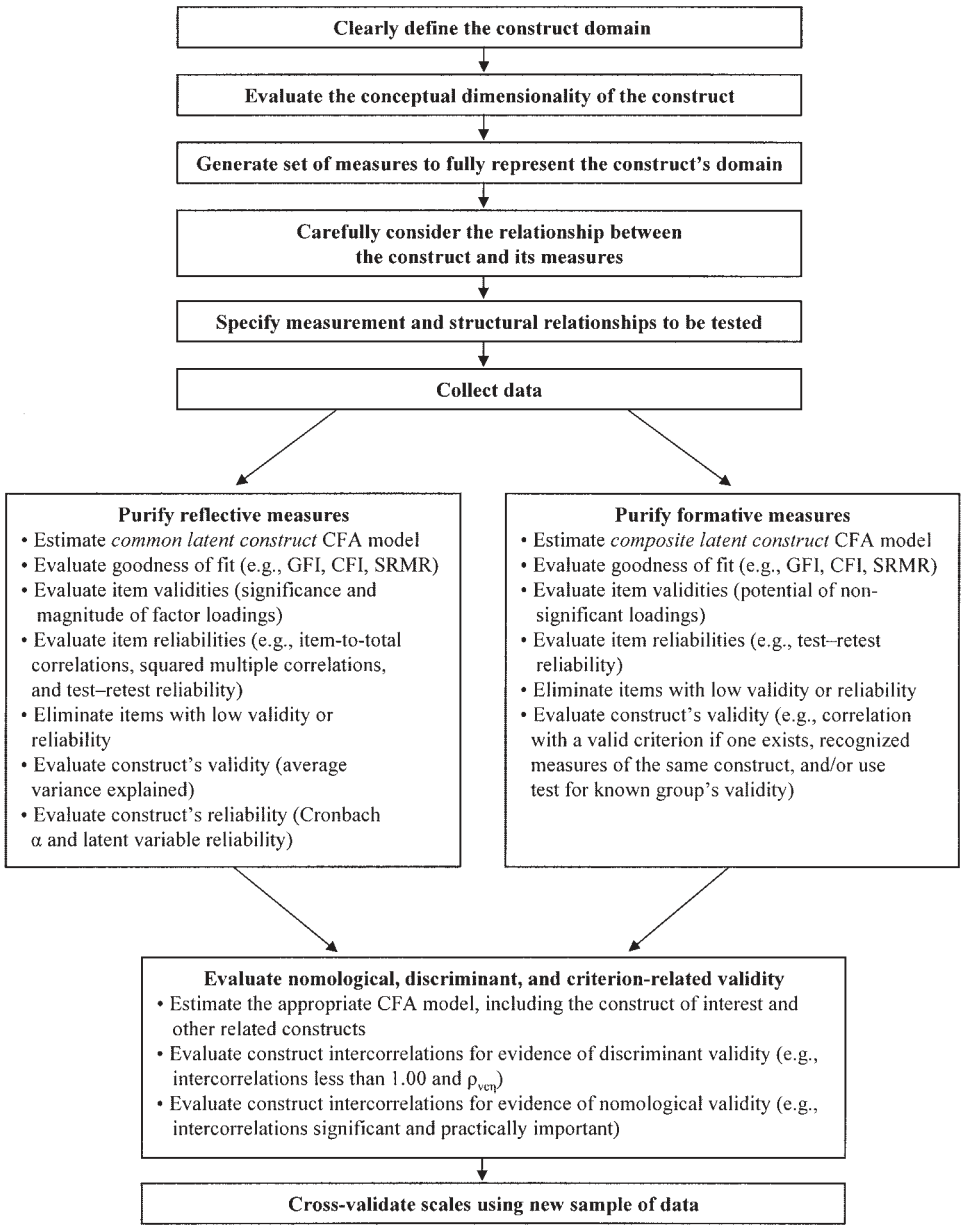

Figure 7. Comparison of scale development processes for reflective and formative constructs.

As indicated in Figure 7, the initial stages involve defining the conceptual domain and evaluating the conceptual dimensionality of the hypothetical construct. Our own review of the literature suggests that the failure to think carefully about the dimensionality of constructs is one of the primary causes of measurement model misspecification. As shown in Figures 2 and 3, constructs are often multidimensional, and their subdimensions may have either formative or reflective measures and may themselves be either formative or reflective indicators of a second-order construct.

The next step is to generate a set of items that completely captures the conceptual domain of the construct (i.e., is not deficient) without being contaminated by other related constructs. For reflective measures, the goal is to develop a representative sample of measures for the construct (cf. Nunnally & Bernstein, 1994; Schwab, 1980), whereas for formative measures, the goal is to develop a set of measures that represents a census of the key elements of the conceptual domain (Bollen & Lennox, 1991).

Once the measures have been generated, the next step is to determine whether the measures are reflective or formative indicators of the construct(s) of interest. This decision can be based on the criteria discussed earlier in this article. Generally speaking, the indicators are *formative* if the following conditions prevail: (a) The indicators are viewed as defining characteristics of the construct, (b) changes in the indicators are expected to explain changes in the construct, (c) the indicators do not necessarily share a common theme, (d) eliminating an indicator may alter the conceptual domain of the construct, and (e) the indicators are not necessarily expected to have the same antecedents and consequences.

Regardless of whether the indicators are formative or reflective, the next step is to establish the scale of measurement for the construct. For reflective-indicator models, this can be done by fixing a path from the latent construct to one of its indicators at 1 or by fixing the variance of the construct at 1. Either of these solutions is acceptable. For formative-indicator models, the scale of measurement is set by either fixing a path from one of the indicators to the composite latent construct at 1 or by fixing the variance of the construct at 1. Once again, either of these solutions is acceptable.

Once this is done, reflective-indicator models should be estimable. However, an additional step is needed to achieve identification for formative-indicator models. To resolve the indeterminacy associated with the construct-level error term, each construct with formative indicators must emit paths to at least two unrelated reflective indicators. As shown in Figure 8, this condition can be satisfied if the construct emits paths to at least two unrelated latent constructs with reflective indicators (see Figure 8A), one reflective indicator and one latent construct with reflective indicators (see Figure 8B), and/or at least two theoretically appropriate reflective indicators (see Figure 8C).

As an illustration of these methods of identification, assume that a researcher is interested in job satisfaction and wishes to model it as a composite latent construct with multiple formative indicators (e.g., satisfaction with pay, satisfaction with coworkers, and satisfaction with the supervisor). In isolation, without any paths entering it or emanating from it, this construct would not be identified. However, one way that it would be identified is if the theoretical structure being tested included paths from job satisfaction to two unrelated latent constructs with reflective indicators, as shown in Figure 8A. Intent to leave the organization and employee voice might be suitable for this purpose, assuming that these constructs had reflective indicators and that they were not causally related. The resulting model would be identified, including the error terms *D1*, *D2*, and *D3*. As shown in Figure 8B, in a situation in which there is only one path emanating from job satisfaction to a latent construct with reflective indicators, another way of achieving identification would be to add a single reflective indicator (V4) of job satisfaction. An example of a potentially appropriate item might be "Overall, how satisfied are you with your job?" Because this new indicator captures overall job satisfaction rather than any one of its individual facets, it is reflective in nature. Therefore, with the addition of this reflective indicator, the job satisfaction construct would now have two paths emanating from it and would be identified (including *D1* and *D2*).

A potentially more versatile method of achieving identification (illustrated in Figure 8C) would be to include two reflective indicators (V4 and V5) for the job satisfaction construct (e.g., "Overall, how satisfied are you with your job?" and "Generally speaking, I am very satisfied with all facets of my job"). The advantage of having two reflective indicators is that it allows (a) a confirmatory factor model to be estimated and (b) the job satis-

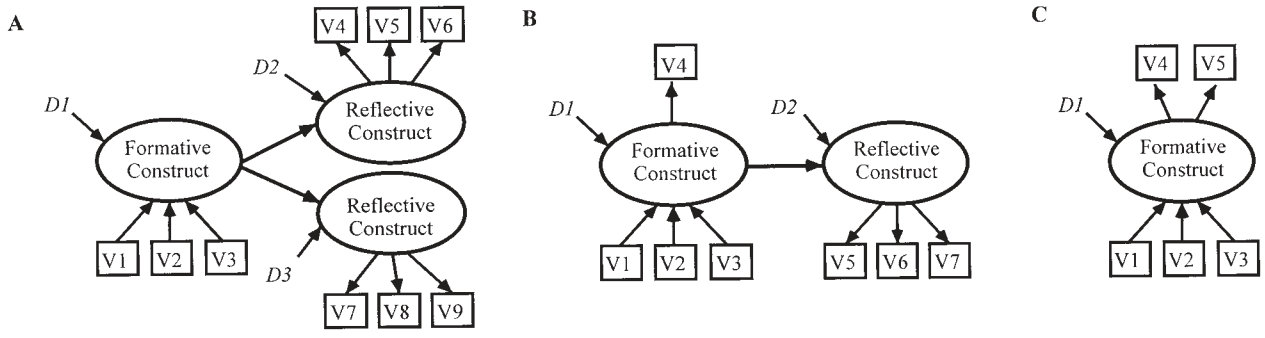

Figure 8. Alternative methods of achieving identification in formative-indicator measurement models. A: Achieving identification by emitting paths to two unrelated latent constructs with reflective indicators. B: Achieving identification by emitting one path to a reflective indicator and one path to a latent construct with reflective indicators. C: Achieving identification by emitting two paths to two theoretically appropriate reflective indicators.

faction construct to occupy any position in a structural model (i.e., endogenous or exogenous). In addition, it is important to note that the measurement model in Figure 8C should be viewed as a single latent construct with a mixture of formative and reflective indicators rather than as a single reflective-indicator latent construct with multiple causes. This is a subtle but important conceptual distinction that arises from the nature of the indicators. It makes sense to interpret the structure in this way because the indicators (whether reflective or formative in nature) all relate to the same conceptual domain specified in the construct definition and are all content-valid operationalizations of the same construct. Indeed, this model is analogous to the redundancy analysis model often used to assess the validity of formative indicators in partial least squares analyses (cf. Chin, 1998).

A final decision that needs to be made when specifying models that include constructs with formative indicators is how to handle the covariances among the exogenous variables and constructs. Two ways have been proposed (cf. MacCallum & Browne, 1993): (a) Constrain all of the covariances among the exogenous latent constructs and manifest variables to be equal to 0, or (b) follow standard practice and estimate the covariances among all exogenous latent constructs and manifest variables. The advantage of the former approach is that model parsimony is not undermined by the addition of a potentially large number of nonhypothesized paths, and as a result, the goodness-of-fit indices will be a better reflection of the veracity of the hypothesized relationships. The advantage of the latter is that the overall fit of the model is not unnecessarily penalized for the covariances among the exogenous variables that are due to factors outside of the model. Of these two approaches, we agree with MacCallum and Browne (1993) that freeing up all of the covariances is a better solution, because it would be a very strong theoretical statement to assume that all of the exogenous variables are perfectly uncorrelated. Indeed, if these covariances were all fixed at 0, it would mean that any common causes of these variables that were outside of the system of relationships represented in the model would contribute to the lack of fit of the proposed model. For this reason, fixing the covariances is not an acceptable solution to this problem.

Once the measurement model has been formally specified, the next step is to collect data for the purposes of evaluating and purifying the measures. The first step in purifying the measures is to estimate a confirmatory factor model and evaluate whether (a) the individual hypothesized measurement relationships are statistically significant, (b) the solution is proper, and (c) the relationships (as a group) are consistent with the sample data. Assuming that these conditions are met, the validity and reliability of the individual items and constructs can be evaluated. For reflective indicators, item validity is equal to $\lambda_{std.}^2$, and item reliability is equal to the squared multiple correlation coefficient when the specific variance is equal to 0 (Bollen, 1989). These values are the same when only one construct influences each measure, as in most confirmatory factor models. For formative indicators, item validity is reflected in the significance and strength of the path from the indicator to the composite latent construct. Multicollinearity among the indicators poses the same problems in this case as it does in multiple regression models, because the relationships between the indicators and the composite latent constructs are analogous to a multiple regression in which the construct is the dependent variable and the indicators are the independent variables. In addition, it is important to note that it is not conceptually or empirically possible for one formative indicator to account for 100% of the variance in the latent construct. It is not conceptually possible for a formative indicator to perfectly represent a composite latent construct, because the construct is defined as being a function of multiple distinct components or parts, so one indicator cannot validly represent the entire conceptual domain. Moreover, it is not empirically possible for one indicator to account for all of the variance in the composite latent construct unless the indicators are perfectly correlated (which is unlikely to be the case). The more indicators there are, the lower the average variance accounted for. Although we are not aware of any absolute standards for judging the absolute magnitudes of the item validities for formative indicators, the relative magnitudes can be compared using the standardized path coefficient.

Another possibility for assessing item validity for formative indicators can be applied if the composite latent construct has at least one reflective indicator that measures the overall composite latent construct. This method uses the reflective indicator in much the same way that a criterion measure would be used to establish criterion-related validity. The validity estimates for the formative indicators of a composite latent construct are obtained by requesting the standardized indirect effects from the structural equation modeling program. This indirect effect is the estimate of the impact of the formative indicator on the composite latent construct multiplied by the estimate of the impact of the composite latent construct on the reflective indicator. In Figure 8B, this would be represented by the indirect effect of V1 on V4. If more than one reflective indicator is available for the construct, the indirect effects of a given formative indicator on each of the reflective indicators could be averaged to provide a more robust estimate of the indicator's validity. This indirect effect represents the relationship between the formative indicator and an overall measure of the construct, and it can be viewed as an index of item validity if one can assume that the overall measure is a valid criterion measure on the basis of its content. Obviously, this technique puts a premium on the content validity of the reflective indicators. If they are invalid, then they cannot be used to provide evidence for item validity.

For a reflective indicator, the item reliability is equal to the item validity or the squared multiple correlation for the item as long as only one latent construct causes each measure. Because item reliability places an upper bound on item validity, it is desirable for the squared multiple correlation for each item to be greater than .50. However, for formative indicators, it is difficult to obtain an estimate of item reliability. There is no squared multiple correlation coefficient for the item because the measurement model posits that the item influences the latent construct rather than vice versa. Consequently, reliability must be assessed through other means. One possibility is to use a test–retest procedure, but this would only be an appropriate index of reliability if the item is expected to be stable over time (cf. Nunnally & Bernstein, 1994, p. 254). Interrater agreement can also be used as an index of item reliability for some types of constructs (e.g., observable performance behaviors). Still another possibility is to correlate each individual item with an alternative measure of the same specific aspect of the construct's domain in a pretest and to use the correlation as an index of reliability. Although not all of these methods can be used for every measure, hopefully at least one will be appropriate.

To purify the measures, one should eliminate items that fail to meet the desired standards for validity and reliability discussed above. However, in doing so, a few caveats are important. First, if the construct-level validity and reliability are good (see below), do not worry if a few of the individual-item reliabilities or validities do not meet the desired standards. Only items with unacceptably low validity or reliability should be eliminated (e.g., Hinkin, 1995; Nunnally & Bernstein, 1994; Spector, 1992). Second, although eliminating the only item that taps an essential aspect of the construct domain is always potentially problematic, this is especially important to keep in mind for composite latent constructs, because the formative indicators must provide a census of the conceptual domain, not just a sample of it (cf. Bollen & Lennox, 1991).

Once item validity and reliability have been assessed, the next step is to evaluate construct-level validity. For constructs with reflective indicators, *convergent validity* can be assessed by the average variance in the items accounted for by the latent construct they represent (cf. Fornell & Larcker, 1981). This can be calculated by averaging the squared multiple correlations (or the squared completely standardized loadings—λ^2s) for the construct's measures. Fornell and Larcker (1981) have argued that for a construct to possess convergent validity, the majority of the variance in its items (i.e., more than 50%) should be accounted for by the underlying construct rather than by measurement error. For constructs with formative indicators, convergent validity at the item level is not relevant, because the composite latent construct model does not imply that the measures should necessarily be correlated. Instead, assessments of construct validity should be based on nomological or criterion-related validity (see below).

For constructs with reflective measures, construct-level reliability can be assessed with an estimate of internal consistency (e.g., Cronbach's alpha) or the somewhat different index of reliability proposed by Bagozzi (1980, p. 181). Generally speaking, the accepted standard for these indices is .70 or above (Nunnally & Bernstein, 1994). For formative-indicator constructs, the concept of internal consistency is not appropriate as a measure of reliability because the indicators are not assumed to be reflections of an underlying latent variable. Indeed, as noted by Bollen and Lennox (1991), formative indicators may be negatively correlated, positively correlated, or completely uncorrelated with each other. Consequently, Cronbach's alpha and Bagozzi's index should not be used to assess reliability and, if applied, may result in the omission of indicators that are essential to the domain of the construct.

One way to assess *discriminant validity* that works for both formative and reflective measures is to test whether the constructs are less than perfectly correlated. This test requires the scale of measurement for each latent construct to be set by fixing its variance at 1, and it should be performed for one pair of constructs at a time. One could also test whether the construct intercorrelation is less than .71. This would test whether the constructs have significantly less than half of their variance in common. Both of these tests can be done by examining the confidence interval around the estimate. A more stringent test that can be used for reflective-indicator models is to examine whether the average variance extracted for each construct ($\rho_{vc\eta}$) is greater than the square of the correlation between the constructs (cf. Fornell & Larcker, 1981). Conceptually, this test requires that each latent construct account for more of the variance in its own indicators than it shares with another construct. This test should be performed for one pair of constructs at a time by averaging the squared multiple correlations for each construct's indicators (separately) and comparing these values to the square of the intercorrelation between the two constructs of interest.

Nomological validity can be assessed using the same procedure, regardless of whether constructs have formative or reflective indicators. Its assessment entails estimating the latent constructs and testing whether their intercorrelations with hypothesized antecedents, correlates, and consequences are significantly greater than 0. Of course, for consistency, the magnitude of the correlations used to establish nomological validity should be greater than the magnitude of the correlations used to establish discriminant validity. This type of validity can also be assessed by using groups with recognized differences on the construct of interest and testing whether the mean level of the construct differs across these groups in the hypothesized direction. For example, if one is interested in assessing the validity of the measures of a quality-of-life construct, one could compare the scores of a group of people with known quality-of-life deficits (e.g., chronic or extended illnesses) with those of another group that does not possess these quality-of-life deficits to see if the scores of the two groups differ in the expected direction.

The final step in the scale development process is to cross-validate the psychometric properties of the scale. This is particularly important if model modifications were made in the scale development and refinement process. Regardless of whether the measures are formative or reflective, either the multigroup approach discussed by Steenkamp and Baumgartner (1998) or the statistical technique developed by Browne and Cudeck (1989, 1993) could be used to cross-validate the scale properties.

Conclusions

In conclusion, this research has shown that there are important distinctions between formative- and reflective-indicator measurement models. Several of the most commonly researched constructs in the field have formative measures that are incorrectly modeled as though they were reflective measures. This is a problem, because as demonstrated by our Monte Carlo simulation, measurement model misspecification can inflate unstandardized structural parameter estimates by as much as 400% or deflate them by as much as 80% and lead to either Type I or Type II errors of inference, depending on whether the endogenous or the exogenous construct is misspecified. Moreover, the simulation results suggest that there is a substantial probability that measurement model misspecification will not be detected with many of the most commonly used goodness-of-fit indices, with the exception of the RMSEA.

This research has several important implications. First, it suggests that some of the empirical findings reported in the literature may be misleading. Even our cursory review of the literature suggests that measurement model misspecification is not uncommon, and because this misspecification can lead to Type I and Type II errors, we have reason to question the validity of the findings of studies that failed to correctly model measurement relationships. Thus, it appears that as Schwab (1980) predicted, the failure to give adequate attention to construct validity and measurement model specification has probably led to substantive con-

clusions about the relationships between constructs that are unwarranted. Although the extent to which this is true is difficult to evaluate without access to the actual data from studies in which the measurement models have been misspecified, in our opinion, this could be a fairly serious problem for the field.

A second implication is that it is essential for researchers to think carefully about the relationships between constructs and their indicators and to make sure that these relationships are correctly modeled. Indeed, our findings would suggest that a researcher's implicit hypotheses about measurement relationships are as important as his or her hypotheses about structural relationships and should be tested empirically. To help with this, we have provided a set of criteria that can be used for deciding on the appropriate measurement model, and we have discussed the specification of constructs with formative indicators in some detail.

A final implication of our research that flows directly from the previous work of Bollen and Lennox (1991) is that it needs to be recognized that some of the procedures for developing and evaluating constructs with reflective indicators cannot be used for constructs with formative indicators. Most of the recommendations for how to develop and evaluate measures are based on classical test theory and its assumption that the measures reflect the underlying constructs they are intended to represent. However, this assumption is not appropriate for formative measures. Therefore, we have provided a set of guidelines for developing, evaluating, and validating constructs with formative indicators, and we have contrasted these recommendations with those for reflective indicators. Although additional refinements will undoubtedly have to be made to our recommendations, we nevertheless believe that they represent a good first step in the development of a set of procedures that researchers can use for constructs with formative measures.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14,* 496–515.

Bacharach, S. B., Bamberger, P. A., & Sonnenstuhl, W. J. (2002). Driven to drink: Managerial control, work-related risk factors, and employee problem drinking. *Academy of Management Journal, 45,* 637–658.

Bagozzi, R. P. (1980). *Causal models in marketing.* New York: Wiley.

Bass, B. M. (1985). *Leadership and performance beyond expectations.* New York: Free Press.

Bass, B. M. (1998). *Transformational leadership: Industrial, military, and educational impact.* Mahwah, NJ: Erlbaum.

Baum, J. R., Locke, E. A., & Smith, K. G. (2001). A multidimensional model of venture growth. *Academy of Management Journal, 44,* 292–303.

Bearden, W. D., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research, 24,* 283–291.

Blalock, H. M., Jr. (1964). *Causal inferences in nonexperimental research.* New York: Norton.

Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality and Quantity, 18,* 377–385.

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

Bollen, K. A. (1990). Overall fit in convariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

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.

Bycio, P., Hackett, R. D., & Allen, J. S. (1995). Further assessments of Bass's (1985) conceptualization of transactional and transformational leadership. *Journal of Applied Psychology, 80,* 468–478.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah, NJ: Erlbaum.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16,* 64–73.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86,* 386–400.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field studies.* Boston: Houghton Mifflin.

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands–resources model of burnout. *Journal of Applied Psychology, 86,* 499–512.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18,* 39–50.

Geyer, A. L. J., & Steyrer, J. M. (1998). Transformational leadership and objective performance in banks. *Applied Psychology: An International Review, 47,* 397–420.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21,* 967–988.

Holtom, B. C., Lee, T. W., & Tidd, S. T. (2002). The relationship between work status congruence and work-related attitudes and behaviors. *Journal of Applied Psychology, 87,* 903–915.

Hom, P. W., & Kinicki, A. J. (2001). Toward and greater understanding of how dissatisfaction drives employee turnover. *Academy of Management Journal, 44,* 975–987.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Law, K. S., & Wong, C. S. (1999). Multidimensional constructs in structural equation analysis: An illustration using the job perception and job satisfaction constructs. *Journal of Management, 25,* 143–160.

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114,* 533–541.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95–107.

MacKenzie, S. B., Podsakoff, P. M., & Ahearne, M. (1998). Antecedents

and consequences of in-role and extra-role performance. *Journal of Marketing, 62, 87–98.*

Masterson, S. S. (2001). A trickle-down model of organizational justice: Relating employees' and customers' perceptions of and reactions to fairness. *Journal of Applied Psychology, 86,* 594–604.

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal, 44,* 1102–1121.

Moorman, R. H. (1991). Relationship between organizational justice and organizational citizenship behaviors: Do fairness perceptions influence employee citizenship? *Journal of Applied Psychology, 76,* 845–855.

Motowidlo, S. J. (2000). Some basic issues related to contextual performance and organizational citizenship behavior in human behavior. *Human Resource Management Review, 10,* 115–126.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Lexington Books.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R., & Fetter, R. (1990). The impact of transformational leader behaviors on employee trust, satisfaction, and organizational citizenship behaviors. *Leadership Quarterly, 1,* 107–142.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal, 4,* 1248–1264.

Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in Organizational Behavior* (Vol. 2, pp. 3–43). Greenwich, CT: JAI Press.

Spector, P. E. (1992). *Summated rating scales: An introduction.* Newbury Park, CA: Sage.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–90.

Tracey, J. B., & Hinkin, T. R. (1998). Transformational leadership or effective managerial practices? *Group and Organization Management, 23,* 220–236.

Wong, C. S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *Leadership Quarterly, 13,* 243–274.

Received July 30, 2003
Revision received July 28, 2004
Accepted August 24, 2004 ■

Low Publication Prices for APA Members and Affiliates

Keeping you up-to-date. All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor.* High School Teacher and International Affiliates receive subscriptions to the *APA Monitor,* and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

Essential resources. APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association,* and on dozens of new topical books each year.

Other benefits of membership. Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

More information. Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.