A Comparison of Approaches to Forming Composite Measures in Structural Equation Models

Ronald S. Landis, Daniel J. Beal and Paul E. Tesluk

Organizational Research Methods 2000 3: 186
DOI: 10.1177/109442810032003

A Comparison of Approaches to Forming Composite Measures in Structural Equation Models

RONALD S. LANDIS
DANIEL J. BEAL
Tulane University

PAUL E. TESLUK
University of Maryland

A common practice in applications of structural equation modeling techniques is to create composite measures from individual items. The purpose of this article was to provide an empirical comparison of several composite formation methods on model fit. Data from 1,177 public school teachers were used to test a model of union commitment in which alternative composite formation methods were used to specify the measurement components of the model. Bootstrapping procedures were used to generate data for two additional sample sizes. Results indicated that the use of composites, in general, resulted in improved overall model fit as compared to treating all items as individual indicators. Lambda values and explained criterion variance indicated that this improved model fit was due to the creation of strong measurement models. Implications of these results for researchers using composites are discussed.

Structural equation modeling (SEM) has become one of the most widely applied data analytic techniques in organizational research. One of the primary reasons for this widespread use is the ability of the technique to assess simultaneously the fit of measurement models and structural models. *Measurement models* test relationships (i.e., paths) between the measures (i.e., manifest variables) and the constructs (i.e., latent variables) that they represent, whereas *structural models* specify relationships between the latent variables of interest.

Authors' Note: The authors would like to thank the editor and two anonymous reviewers for their constructive reviews. In addition, the authors would like to thank Jeff Edwards for his helpful comments on an earlier version of this article. Portions of this research were presented as part of a symposium (P. Tesluk, Chair) at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, Georgia, April 1999. Please address correspondence concerning this article to Ronald S. Landis, Department of Psychology, Tulane University, New Orleans, LA 70118; e-mail: rlandis1@mailhost.tcs.tulane.edu.

Organizational Research Methods, Vol. 3 No. 2, April 2000 186-207

Despite the substantial attention paid to justifying and specifying the relationships between latent variables in many applications of SEM, relatively little attention has been focused on the relationships incorporated in measurement models (Bagozzi & Edwards, 1998). This is not surprising considering that it is the structural relationships that represent the primary study hypotheses. However, because questions of construct validity are equally important and directly affect the substantive models being tested, measurement models deserve more attention (Bagozzi & Edwards, 1998).

Bagozzi and Edwards (1998) provided a conceptual framework for representing constructs in which they presented methods of construct specification using confirmatory factor analytic methods. Arguing that issues related to construct depth and dimensionality should determine the specificity of the relationships, four alternative measurement models were described. The *total disaggregation model* is characterized by the treatment of all relevant items as indicators of the latent construct of interest. The *partial disaggregation model* and the *partial aggregation model* involve the combination (e.g., through summing or averaging) of items into subsets, which, in turn, are treated as indicators of the latent construct. Finally, the combination of all items in a particular scale into a single indicator of the latent construct is termed the *total aggregation model*.

Although previous authors have not used the terminology suggested by Bagozzi and Edwards (1998), they have adopted strategies for creating composites (also referred to as *testlets* or *item parcels*) that can be described by this nomenclature (e.g., Mathieu & Farr, 1991; Mathieu, Tannenbaum, & Salas, 1992). Unlike the theoretically grounded models presented by Bagozzi and Edwards, however, the rationale for creating composites in practice has been driven by more practical concerns. Including all items as individual indicators in a full SEM analysis requires a substantially larger sample size as the number of indicators increases. Because many studies do not have the required sample sizes for these analyses, researchers often adopt composite formation techniques to reduce the number of estimated parameters in the tested model. Thus, to produce more stable estimates of structural relationships, researchers have sacrificed testing total disaggregation models in favor of partial disaggregation, partial aggregation, and total aggregation models.

Because several alternative composite formation methods have been applied in the extant literature, the purpose of this article is to provide an empirical comparison of these methods. The following section presents a review of the most commonly applied methods.

A Review of Composite Formation in SEM

As previously stated, the purpose in many applications of SEM is to describe both the structural and measurement relationships of a specified model. Indeed, one of the chief benefits of using SEM techniques is that it allows for this concurrent assessment of reliability and validity. Unfortunately, to assess the fit of the full model, the number of cases must be significantly larger than the number of parameters estimated. Although there is no single criterion with regard to the necessary sample size, several perspectives have been offered. Anderson and Gerbing (1988) stated that a minimum required sample size was 150, whereas Kelloway (1998) suggested that at least 200 observations represented an appropriate minimum. Alternatively, Bentler and Chou

(1987) framed the issue in a slightly different way and suggested that the ratio of sample size to estimated parameters be between 5:1 and 10:1. In situations in which just a few items are used as indicators of each latent construct, the Bentler and Chou criterion can be met with a relatively modest sample size. However, as the number of items per construct increases, meeting this criterion becomes less likely given the number of cases typically found in psychological and organizational research.

An alternative perspective with regard to the sample size required for accurate model testing has recently been advocated by Marsh, Hau, Balla, and Grayson (1998). Based on simulation work, Marsh et al. argued that traditional rules of thumb with regard to the ratio of sample size to estimated parameters might be inappropriate and that researchers should consider using more indicators than is evident in current practice. Although the recommendations of Marsh et al. may be justified given their results, the authors themselves suggested that several features of the simulations might limit the generalizability of their conclusions. In a section specifically dealing with item parcels, Marsh et al. noted that their study evaluated only a very limited range of relevant variables. They further suggested that the use of nonnormal data, more complicated data structures, and various levels of misfit, among others, would greatly complicate studies of the effectiveness of item parcels but "might also show some advantages of item parcels over items—particularly when *N* is small" (p. 217). In short, the impact of using composite measures in models with measurement and structural components and with real world data remains an important unanswered question.

The preceding discussion points to a tension between testing structural models by incorporating finer grained (i.e., total disaggregation) measurement models versus coarser grained (i.e., total aggregation) measurement models. The benefit of the total disaggregation model is the ability to evaluate the performance of each item in a scale. Despite the desire to make use of this richer information, commonly encountered sample sizes often result in the failure to obtain stable parameter estimates. In light of this and to obtain an accurate assessment of structural relationships, researchers often have little choice but to collapse the items in a particular scale into multi-item composites. There is no widely agreed on method for forming multi-item composites, although it is a fairly common practice and a variety of methods have been described. In fact, no less than six methods appear to be potentially useful for accomplishing this purpose.

One of the most frequently reported methods (Brooke, Russell, & Price, 1988; Lent, Lopez, Brown, & Gore, 1996; Mathieu & Farr, 1991; Mossholder, Settoon, Harris, & Armenakis, 1995) is completed by subjecting all items on a particular scale to a factor analysis in which a single-factor solution is specified (hereafter referred to as the single-factor [SFA] method). The resulting factor loadings are examined, and pairing the item with the highest loading with the item having the lowest loading forms the first composite. Pairing the items with the remaining highest and lowest loadings then forms the second composite. This process is repeated until all items have been assigned to composites. The underlying purpose of this method is to distill the original set of scale items to a reduced number of indicators that are empirically balanced measures of the construct (Mathieu & Farr, 1991).

An alternative technique described by Cramer (1996) produces composites based on item intercorrelations (hereafter referred to as the correlational [R] method). First, bivariate relationships between all scale items are computed. Second, those items with the strongest correlation are paired to form the first composite. Next, the two remain-

ing items with the strongest correlation are paired, and this continues until all items have been assigned. The intent of this approach, like the previous one, is to create composites that are empirically similar to one another.

Both of the preceding techniques require an intermediate analysis (i.e., factor analysis or examination of the correlation matrix) prior to composite formation. Alternatively, if all items on a scale are equivalent measures of a particular construct, random assignment of items to composites should produce empirically balanced measures (i.e., equal means, variances, and reliabilities) as well. Unlike the first two techniques, however, the random assignment of items to composites does not require the completion of an intermediate step. Williams and Anderson (1994) described an application of this method (hereafter referred to as the random [RAND] method). To the extent that items are not equivalent measures of the focal construct, this technique will most likely produce nonequivalent composites.

Although the previous methods have been used in the literature, there are at least three other methods that could also prove useful. One common characteristic of the approaches described in the preceding section is a reliance on empirical relationships between items. Another approach to developing composites would involve employing a content-oriented strategy in which items are assigned to composites based on existing theory or rational judgment (hereafter referred to as the content [CONT] method). Many unidimensional scales include subsets of items measuring unique dimensions of a broader construct of interest. If this is the case, a content-oriented approach makes logical sense as a means of composite formation. In fact, a major criticism of the empirical approaches might be that using known relationships between items to make decisions about composites loads the dice in terms of developing strong measurement models, which, in turn, could lead to overestimated structural models.

In addition to rationally grouping items, a related method would be to perform an exploratory factor analysis on each scale and assign items based on their factor loadings (hereafter referred to as the exploratory factor analysis method). Unlike the factor analytic method previously described (i.e., SFA), this approach would let the chips fall as they may in terms of creating composites. If a two-factor solution fits a given set of items best, it would be preferred. Alternatively, if a three-factor solution provided a better fit, three resulting composites would be used. This approach provides a complementary perspective on the content-oriented strategy previously discussed in that no a priori decisions are made about the number of composites. In contrast to the content-oriented strategy, however, the EFA would represent a data driven approach.

A final strategy not described in the extant literature is to create truly empirically equivalent composite measures (hereafter referred to as the empirically equivalent method). According to this philosophy, items would be assigned to composites such that these measures would have equal means, variances, and reliabilities (Nunnally & Bernstein, 1994, p. 223). This approach is similar to the empirically driven models described earlier except that in addition to factor-loading information, means, standard deviations, and reliabilities are used to create the composites. If the measurement models become stronger as composites approach empirical equivalence, this technique should yield the greatest improvement in overall model fit.

A summary of the previously described methods of composite formation is included as Table 1.

Table 1
Summary of Composite Formation Methods

Method	*Description*
Single factor (SFA)	Pair off items with highest and lowest loadings as first composite based on a single-factor solution; continue pairing until items are exhausted
Correlational (R)	Pair off items with highest intercorrelation as first composite; continue pairing until items are exhausted
Random (RAND)	Randomly assign items to composites
Content (CONT)	Create composites based on rational grouping(s) of items
Exploratory factor analysis (EFA)	Create composites based on results from exploratory factor analysis
Empirically equivalent (EE)	Create composites with equal means, variances, and reliabilities

Conceptual Considerations and Composite Formation

In addition to the practical issues just discussed, there are important theoretical considerations related to forming multi-item composites. For the most part, composites are created from theoretically unidimensional scales. To the extent that all items tap only one underlying construct, there would be only minor differences expected as a result of the various composite formation strategies. If the items measure more than one factor, however, then composite formation is likely to be affected (Hall, Snell, & Foust, 1999). Hall et al. (1999) described how the presence of an unmodeled secondary factor could substantially influence the parameter estimates associated with developed composites. These influences would be due to the misspecification of the underlying measurement model. Hall et al. further argued that when items share a secondary factor, the best choice of composite formation is one in which these items are allocated to the same composite(s) as opposed to distributed across all composites.

One way to view the practice of forming composites is that it is essentially the same as the practice of developing multi-item measures from scratch. Within classical test theory (CTT), the domain-sampling model is often used to guide test construction (Nunnally & Bernstein, 1994). The domain-sampling model is based on the notion that measurement error can be best construed in terms of responses to a set of items randomly sampled from a universe of possible items. Although items are not truly sampled but composed, the model generally works well because the items on most tests are sufficiently broad and the results are similar to those that are expected had they been sampled from the universe of possible items. Furthermore, there is no requirement with regard to the number of specific items that are sampled, and the model can be developed without regard to item type.

In situations in which a set of items truly measures a single underlying factor, the choice of composite formation strategy should not substantially influence model fit. To the extent that a given set of items taps multiple factors, however, the choice of composite formation would be expected to influence the overall model fit by impacting the measurement model. Following the logic of the domain-sampling model, if one was interested in measuring a construct with multiple dimensions, the resulting set of items should provide an adequate representation of the underlying construct dimensionality.

For example, if the underlying construct was composed of two equal subdimensions, the resulting measure should include equal numbers of items related to these dimensions. To the extent that one dimension was overrepresented, reliability would be expected to suffer.

In the case of composite formation, the researcher has already defined the universe of items (i.e., those that were written and administered). As a result, it would be expected that composite formation strategies that produced individual composites measuring the construct space in a representative manner would produce better fitting measurement models. An example may prove useful in illustrating this point. Consider an eight-item scale designed to measure job satisfaction. At the time of scale construction, the researcher defined *job satisfaction* as a unidimensional construct but included specific items that measured two aspects of the construct: satisfaction with supervision and satisfaction with extrinsic rewards. Furthermore, assume that there were four questions for each subdimension. Due to sample size limitations, the researcher decided to use two composites in constructing the measurement model related to this construct. Based on the logic of the domain-sampling model, each of the two composites should most closely approximate the underlying dimensionality of the construct. In short, each four-item composite should include two items from each of the two subdimensions (i.e., satisfaction with supervision and satisfaction with extrinsic rewards) as this breakdown most clearly produces measures that mirror the dimensionality of the initial construct.

This is a different conclusion to the process than if the researcher had started out acknowledging the multidimensional nature of the construct. In such a case, it would have been preferable to construct two four-item scales that each tapped the individual dimension of interest. In the formation of composites, however, this practice appears somewhat illogical in that the universe of items has already been defined. This position is consistent with the work of Hall et al. (1999) in that homogeneous composites make sense when viewed from the perspective of composite construction from the theory down to the composites. In practice, however, composite formation occurs from the items up to the composites. In these cases, it would seem that the more logically consistent and defensible position is to create composites that are as similar to one another as possible. In fact, Little, Lindenberger, and Nesselroade (1999) advance the argument that in situations in which a theory is not strong enough to guide the selection of specific indicators a priori, the selection of indicators should span the domain rather than be highly targeted. Thus, heterogeneous composites would likely be preferable to more homogeneous composites. Little et al. further argued that more focused construct definition is preferable to the selection of items based on convenience. The challenge for the researcher using composites is to recognize that methods that rely more heavily on empirical relationships between items may create the illusion of capturing the true structure of the data.

The previously described methods for forming composites differ in the extent to which items are allocated to composites, which, in turn, could lead to differences in the dimensionality of the composites. First, consider the SFA and the EE approaches. Because items in the SFA technique are paired off using their loadings on a single, underlying factor, there is a strong possibility that to the extent that additional factors (either substantive or methodological) explain the variation between items, these will be distributed across composites. Similarly, the EE method should produce dimensionally similar composites in that an explicit attempt is made to identify similar items

(in terms of means, variances, and reliabilities) and explicitly force them into different composites. This method should result in distributing items that represent the same theoretical dimension across composites. That is, this approach should produce composites that are similar to one another and reproduce the dimensionality of the initial construct as measured.

Alternatively, the EFA and CONT method should produce composites that are dimensionally distinct from one another. These methods attempt to quantitatively (EFA) or qualitatively (CONT) identify specific subdimensions and produce composites based on clustering similar items. As a result, these techniques are more likely to produce composites in the vein suggested by Hall et al. (1999).

The R method is perhaps the least predictable. Using the previous example, assume that the subdimensions (i.e., satisfaction with supervision and satisfaction with extrinsic rewards) are relatively independent from one another. In this case, the intercorrelations within the two four-item sets will be high, but the item intercorrelations across the sets will be low. The first pair of items selected and subsequently placed on the first composite will be measures of the same subdimension. The second two items may or may not be from the same dimension as the first set. Thus, the second composite might have similar or dissimilar items as the first. As the number of items increase, the likelihood of creating conceptually pure composites becomes smaller. As a result, it is more likely that the R method will produce heterogeneous composites as the number of items on the original scale increases. In sum, the R method should produce composites most similar to those produced by the SFA and EE approaches, although in some instances, factorially pure composites may be produced.

Similar to the R method, the RAND method should be somewhat unpredictable, although it is more likely to produce factorially heterogeneous composites. More specifically, on a multidimensional test, there is only one random split of the items that will produce unidimensional composites. Thus, it is more likely that heterogeneous composites will be formed. Again, this suggests that the RAND method will produce composites that are similar to those produced by the SFA and EE methods.

The Study

The purpose of this study was to provide an empirical comparison of the methods previously described. Based on the ideas described by Bagozzi and Edwards (1998) and Hall et al. (1999), it was expected that the various methods would not produce equivalently fitting models. Specifically, methods that produce empirically balanced composites were expected to produce models with better overall fit than methods that do not produce empirically balanced composites. Furthermore, it was anticipated that this better overall model fit would be due, in part, to the superior fit of the measurement models (i.e., the lambda values) associated with the more empirically driven methods.

Method

Participants and Measures

The data were collected as part of a larger study examining the factors that influence union commitment and participation (Agars, Unckless, & Tesluk, 1998). Specifically,

responses to several work attitude scales were collected from unionized public school teachers in a large, northeastern state. A stratified random-sampling strategy was used to recruit participants (N = 4,000) from the union's 446 locals. For this study, 1,177 respondents provided usable data.

Multi-item measures were used to assess five latent constructs. In addition, tenure was measured via a single open-ended item that asked, "How many total years have you been employed as an educator?" Job satisfaction was measured with six items developed by members of the union's research staff. These items asked respondents the extent to which they felt satisfied with student behavior, personal fulfillment, pay, benefits, career advancement, and their job overall. Satisfaction with the union's decision making was assessed using a six-item measure adapted from Hammer and Wazeter (1993).

An eight-item scale adapted from Gordon, Philpot, Burt, Thompson, and Spiller (1980) was used to assess reasons for joining the local union association. Thirteen items taken from Gordon et al. (1980) and Kelloway, Catano, and Southwell (1992) were used to measure union socialization. Finally, union commitment was assessed using the 13-item scale described by Kelloway et al. (1992).

Model Trimming

The hypothesized relationships between the measured and latent variables are illustrated in Figure 1. Although the purpose of this study was to compare various methods for creating composites, not to test specific hypotheses per se, the model was derived from the extant union commitment literature. Based on previous findings in this area (e.g., Barling, Fullagar, & Kelloway, 1992; Barling, Wade, & Fullagar, 1990; Fullagar, Clark, Gallagher, & Gordon, 1994), it was hypothesized that tenure (TENURE), level of job satisfaction (JOBSAT), reasons for joining the union (REASON), socialization experiences in the union (USOC), and satisfaction with the union's decision-making process (DECSAT) would influence the level of commitment felt toward the union (UCOMM).

Anderson and Gerbing (1988) recommended that an original set of items be trimmed before testing the proposed structural relationships of a hypothesized model. In this analysis, relationships between indicators and constructs were examined for significance. Specifically, an item was assigned to a factor when its loading was at least twice its standard error (Anderson & Gerbing, 1988). Using this criterion, an item was dropped if it failed to load on the latent variable for which it was written *or* if it had a significant loading on an irrelevant latent variable. In short, the goal of this method is to retain only those items that measure the construct for which they were intended and to determine the relative independence of the constructs from one another.

After employing the Anderson and Gerbing (1988) procedure with the data, several modifications were made. It should be noted that items discarded based on the model-trimming procedure were dropped due to their significant loadings on multiple latent variables. The job satisfaction measure was reduced to four items (α = .74). The reasons-for-joining-the-union scale was reduced to six items (α = .71), whereas the measures of union socialization and union commitment were both reduced to nine item measures (α = .87 for both). The satisfaction-with-union-decision-making scale retained all original items (α = .89). Item characteristics (e.g., means, standard deviations, skewness, and kurtosis) for the original data and results of EFAs of the original scales are provided in Appendices A and B, respectively.

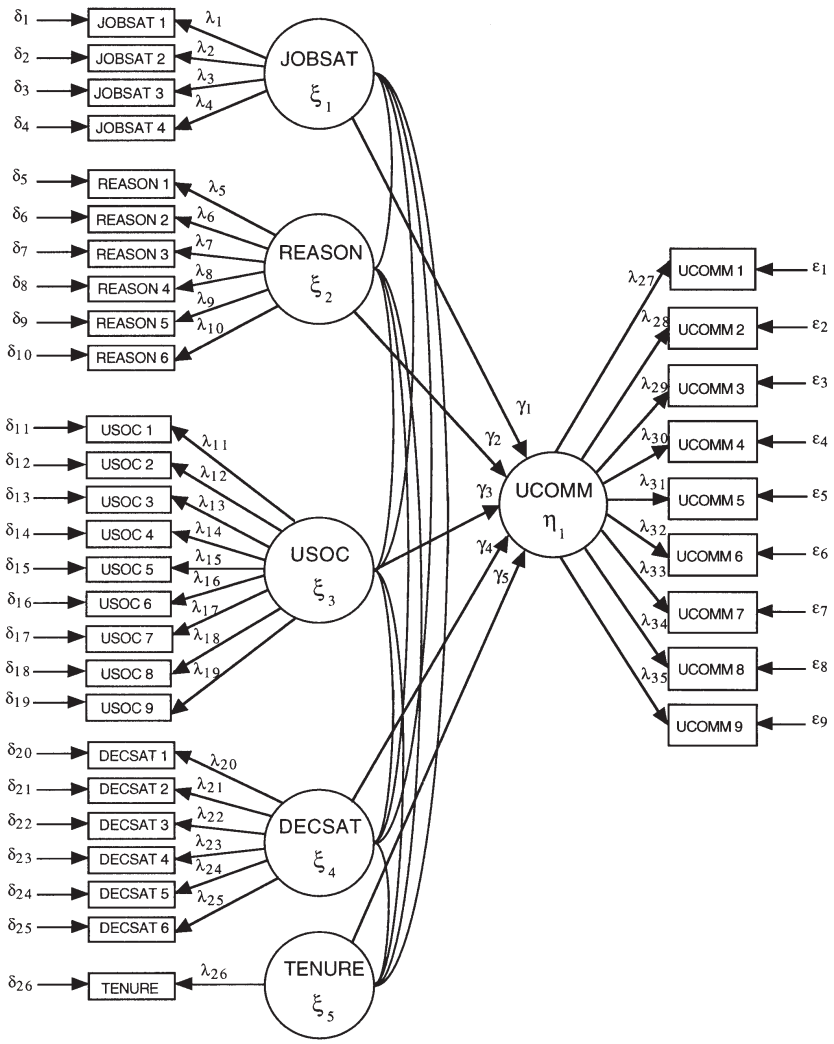

Figure 1: Model of Measurement and Structural Relationships Between Study Variables

Note. TENURE = tenure. JOBSAT = level of job satisfaction. REASON = reasons for joining the union. USOC = socialization experiences in the union. DECSAT = satisfaction with the union's decision-making process. UCOMM = level of commitment felt toward the union.

Data Generation Procedures and Analyses

From the original 1,177 observations, sampling-with-replacement procedures were used to generate 10 data sets for each of three sample sizes: 100, 300, and 1,177.[1] For purposes of comparison with the composite methods, a full model (hereafter referred to as the trimmed [TRIM] method) in which all indicators that met the Ander-

son and Gerbing (1988) criteria were retained was tested at each of the above sample sizes. In addition, each of the previously described methods was used to form composites, and models were tested using EQS for Windows (Bentler, 1995). The structural model was always based on the model included as Figure 1. The only differences across methods were the specified measurement models. It should be noted that for all models tested, the reliability of the tenure measure was assumed to be 1.00. As a result, associated values for lambda and theta-delta for this variable were set to 1.00 and 0.00, respectively. Furthermore, the variances of the latent exogenous variables were always fixed to be 1.0. The following section describes the specific procedures used to generate composites with the data.

For the TRIM model, a correlation matrix of all individual items was generated and used as the EQS input.[2] In addition to the structural relationships, this model estimated loadings for four indicators for the JOBSAT factor, six indicators for the REASON factor, nine indicators for the USOC factor, six indicators for the DECSAT factor, and nine indicators for the UCOMM factor.

The first and third authors completed a content evaluation of the items, and composites in the CONT method were formed based on the apparent dimensionality of the scales. This resulted in three composites for the REASON, USOC, and UCOMM factor; two composites for the DECSAT factor; and one composite for the JOBSAT factor. For each of the 10 samples at each sample size, alpha was computed for the JOBSAT variable, and composite variables were created for the other subscales by summing all of the items within each subscale. Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

For the EFA method, a principal components analysis was performed on each scale for each of the 10 samples at each sample size. Kaiser's criterion and examination of the scree plot were used to identify the factor structure for each scale. If the analysis returned a single-factor solution, a single composite was formed and alpha was used to fix the path between the latent variable and the composite. If a multiple-factor solution was returned, composites were formed by summing the items that loaded onto each factor (i.e., two-factor solutions had two composites whose items were determined by the principal component analysis). Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

In the RAND method, two composites were formed by randomly assigning items to one of the two composites and summing the items within a composite for each scale for each of the 10 samples at each sample size. If there were an odd number of items, the extra item was placed into a composite randomly. Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

Creating composites using the R method involved computing correlations between items within each scale for each of the 10 samples at each level of *N*. For each scale, two composites were formed by placing the two items with the highest correlation in one composite, the two items with the next highest correlation on the other composite, and so forth. If there were an odd number of items, the extra item was placed into the second composite. Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

To create composites using the SFA method, a principal components analysis was run for each scale for each of the 10 samples at all levels of *N*. Unlike in the EFA method, this analysis was constrained to produce a single-factor solution. Using this

information, two composites were then formed for each variable. Specifically, the item with the highest loading and the item with the lowest loading were paired and placed into the first composite. The next highest and lowest loading items were then paired and placed into the second composite. This was repeated until all items were assigned to a composite. If there were an odd number of items, the extra item was placed into the second composite. Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

Finally, composites in the EE method were created by first computing item means, item standard deviations, and item-total correlations for each scale for each of the 10 samples at each level of *N*. Two composites were then created for each variable using this information. Specifically, items with approximately the same mean, standard deviation, and item-total correlation were identified, and one item was placed into one composite and the other item was placed into the second composite. If there were an odd number of items, the extra item was placed into the composite that best equated the two subscales. For example, if one composite's reliability was slightly lower than the other, and the extra item had a high item-total correlation, the extra item was placed in the composite with the lower reliability, thus helping to equate the two composites. Correlations were then generated between every subscale composite in the data, and the resulting matrix was used as the EQS input.

Fit Statistics

Several commonly used indices were used to evaluate the overall model fit. As Medsker, Williams, and Holahan (1994) noted, a variety of fit statistics have been reported in the literature. In addition to chi-square values, we report fit indices that have traditionally been used by researchers and those commonly produced by both the EQS and LISREL (Jöreskog & Sörbom, 1996) programs. One of the fit statistics reported is the nonnormed fit index (NNFI) (Bentler & Bonnett, 1980). Hu and Bentler (1999) have recently suggested that values greater than .95 are desirable for the NNFI. Another model fit index reported in this study is the comparative fit index (CFI). Bentler (1990) argued that the CFI provides a better estimation of model fit than many other alternative indexes. Similar to the NNFI, a cutoff close to .95 is needed before it can be concluded that there is a relatively good model fit (Hu & Bentler, 1999). Finally, the root mean square error of approximation (RMSEA) is reported. The RMSEA is based on the residuals resulting from the discrepancies between the original and the reproduced matrices (Steiger, 1990). Hu and Bentler (1999) suggested .06 as a reasonable cutoff value for the RMSEA.

Results

Overall fit statistics for all models when $N = 1{,}177$ are presented in the top panel of Table 2. Model fit statistics for the remaining samples of 300 and 100 are included in the second and third panels of Table 2, respectively.

Several notable trends are apparent in these results. First, the total disaggregation technique (the TRIM model) displays relatively modest to poor fit across all sample sizes. For example, the CFI values for this model range from .834 to .621. In addition to the relatively poor fit indices associated with the TRIM model, the ratios of sample size to estimated parameters were also the smallest for this approach. These ratios,

Table 2
Mean Model Fit Indices

	Model						
	TRIM	*SFA*	*R*	*EE*	*RAND*	*CONT*	*EFA*
Model with *N* = 1,177							
df	546	30	30	30	30	53	8 to 13
Fit index							
Chi-square	4305.03	162.19	186.49	116.22	169.30	832.42	305.40
CFI	$.834_{e}$	$.981_{a, b}$	$.977_{b}$	$.988_{a}$	$.981_{a, b}$	$.888_{d}$	$.900_{c}$
NNFI	.819	.966	.958	.978	.965	.836	.753
RMSEA	.077	.061	.066	.049	.062	.112	.163
Model with *N* = 300							
df	546	30	30	30	30	53	8 to 12
Fit index							
Chi-square	1593.11	69.57	72.43	59.92	77.19	250.00	77.15
CFI	$.822_{d}$	$.978_{a}$	$.975_{a}$	$.983_{a}$	$.974_{a}$	$.887_{c}$	$.908_{b}$
NNFI	.806	.959	.954	.969	.952	.835	.770
RMSEA	.080	.065	.068	.057	.071	.112	.152
Model with *N* = 100							
df	546	30	30	30	30	53	12 to 25
Fit index							
Chi-square	1595.10	75.50	68.40	59.11	79.99	118.43	54.76
CFI	$.621_{d}$	$.933_{a}$	$.940_{a}$	$.952_{a}$	$.925_{a, b}$	$.888_{b, c}$	$.886_{c}$
NNFI	.587	.878	.891	.966	.863	.835	.763
RMSEA	.139	.122	.109	.097	.127	.111	.147

Note. CFI = comparative fit index. NNFI = nonnormed fit index. RMSEA = root mean square error of approximation. Mean CFIs in the same row that do not share the same subscript differ at $p < .05$ in the Bonferroni least significant difference comparison. TRIM = trimmed. SFA = single factor. R = correlational. EE = empirically equivalent. RAND = random. CONT = content. EFA = exploratory factor analysis.

included in Table 3, ranged from a high of approximately 16:1 when *N* = 1,177 to a low of approximately 1:1 when *N* = 100. The value associated with the largest sample size clearly exceeds the rule of thumb advocated by Bentler and Chou (1987). Given this, it might be tempting to forgo the creation of composites in favor of the total disaggregation model. However, an examination of the overall model fit statistics indicates that this model would not be considered to provide a good explanation of the data. As sample size decreases, fit statistics suggest that this model continues to provide a relatively poor fit to the data, while resulting in ratios of sample size to estimated paths that fail to reach even the lower bound (i.e., 5:1) of Bentler and Chou's rule of thumb.

Overall model fit was consistently superior for the RAND technique and those relying on the empirical creation of parallel composites (i.e., the SFA, R, and EE methods). These methods resulted in ratios of cases to estimated parameters ranging from 44:1 for *N* = 1,177 to 4:1 for *N* = 100. In comparison to the TRIM model, these techniques provide uniformly higher ratios.

Table 2 provides the results of an analysis of the differences of the CFIs for each method at each sample size. Bonferroni least significant difference (LSD) comparisons revealed that the EE, SFA, and RAND methods produced nonsignificant differences in model fit at all three sample sizes. At the largest sample size, the R method also resulted in a fit that was not significantly different from that of the SFA and RAND methods. These four methods (EE, SFA, RAND, and R) were also not significantly

Table 3
Ratios of Sample Size to Estimated Paths for All Models

	Model			
N	*TRIM*	*EMP*	*CONT*	*EFA*
1,177	15.69	43.59	30.97	55.78
300	4.00	11.11	7.89	13.75
100	1.33	3.70	2.63	3.61

Note. Because the correlational, empirically equivalent, single-factor, and random methods produce models with the same number of paths, they have been collapsed into the EMP column. TRIM = trimmed. CONT = content. EFA = exploratory factor analysis.

different at the other two sample sizes. The other methods, however, did result in model fit that was significantly different from the preceding four methods and from each other.

Another important point to note from these results is that they are not completely explained by degrees of freedom. For example, the CONT method produces a model with more degrees of freedom yet results in poorer model fit. Alternatively, the EFA approach produces models with fewer degrees of freedom and also results in poorer values for the fit statistics.

Although the overall model fit statistics appear to tell a consistent story, an examination of the lambda values provides a different perspective on the influence of composites. Table 4 includes the average lambda values associated with the composites of each latent variable across various methods for sample sizes of 1,177, 300, and 100, respectively. It is apparent from these tables that the TRIM method generally produces more poorly fitting measurement models than all other methods. This is not surprising given that the TRIM method treats all items in a scale as individual indicators of the latent variable. An examination of the lambda values associated with those methods that produced the best overall model fit (i.e., the SFA, R, RAND, and EE methods) suggests that these approaches may produce better fit through the creation of more reliable manifest indicators than the TRIM method.

Table 5 contains values of the average criterion variance explained by each of the composite formation methods across all sample sizes. At sample sizes of 1,177 and 300, the TRIM model consistently explains the most variance in the criterion. Unlike the overall model fit values, the average variance explained suggests that the SFA, R, EE, and RAND methods generally are the least appealing alternatives.

The information presented in Table 6 illustrates the effect that each of the composite methods had on the estimates of the structural paths in the model. From this data, it is clear that the TRIM model produces structural paths that were often substantially different from those produced by any of the composite methods. For example, at N = 1,177, the structural path between the JOBSAT latent variable and the UCOMM latent variable in the TRIM model was .477, whereas the composite methods produced paths ranging from –.028 to .013. For the path between DECSAT and UCOMM, the value for the TRIM model was –.001, whereas the composite methods produced paths ranging from .160 to .344. Considering these differences in addition to those associated with the overall percentage of variance accounted for in the UCOMM variable (see Table 5), it is the case that the TRIM and CONT methods result in generally larger

Table 4
Mean Lambda Values for Measures of Each Latent Variable

	Model													
	TRIM		*SFA*		*R*		*EE*		*RAND*		*CONT*		*EFA*	
Latent Variable	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Model with *N* = 1,177														
JOBSAT	.672	.229	.788	.125	.758	.084	.782	.106	.774	.134	.866	.018	.918	.011
REASON	.581	.190	.782	.125	.965	.164	.834	.106	.808	.150	.620	.267	.804	.251
USOC	.682	.058	.900	.047	.882	.077	.881	.018	.898	.055	.795	.079	.994	.001
DECSAT	.763	.056	.897	.021	.887	.024	.896	.021	.906	.019	.883	.029	.995	.001
UCOMM	.696	.139	.911	.052	.910	.068	.925	.043	.916	.052	.664	.228	.794	.191
Model with *N* = 300														
JOBSAT	.670	.238	.785	.124	.759	.143	.773	.164	.775	.122	.860	.044	.916	.023
REASON	.575	.201	.798	.107	.751	.160	.767	.075	.804	.143	.614	.273	.754	.291
USOC	.684	.071	.893	.059	.886	.080	.900	.073	.890	.081	.795	.089	.994	.001
DECSAT	.766	.060	.890	.028	.884	.025	.883	.037	.896	.030	.885	.033	.995	.001
UCOMM	.699	.141	.913	.053	.907	.075	.924	.053	.905	.087	.662	.225	.797	.189
Model with *N* = 100														
JOBSAT	.693	.234	.796	.111	.771	.103	.807	.113	.777	.075	.874	.047	.938	.016
REASON	.579	.263	.789	.133	.817	.140	.812	.152	.795	.202	.611	.271	.675	.302
USOC	.649	.142	.889	.072	.872	.078	.891	.074	.888	.081	.775	.120	.757	.164
DECSAT	.802	.079	.920	.063	.911	.062	.934	.043	.927	.046	.881	.066	.997	.001
UCOMM	.660	.217	.905	.073	.888	.078	.907	.075	.875	.076	.661	.249	.680	.267

Note. TRIM = trimmed. SFA = single factor. R = correlational. EE = empirically equivalent. RAND = random. CONT = content. EFA = exploratory factor analysis. JOBSAT = job satisfaction. REASON = reasons for joining the union. USOC = involvement in the union. DECSAT = satisfaction with the union. UCOMM = commitment to the union.

Table 5
Mean Variance Explained in Union Commitment (UCOMM)

	Model													
	TRIM		*SFA*		*R*		*EE*		*RAND*		*CONT*		*EFA*	
N	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1,177	.777	.017	.679	.025	.690	.020	.651	.021	.674	.030	.746	.020	.609	.031
300	.765	.046	.675	.052	.660	.083	.661	.102	.679	.067	.752	.035	.618	.067
100	.723	.084	.705	.090	.718	.070	.721	.082	.729	.092	.794	.105	.815	.138

Note. TRIM = trimmed. SFA = single factor. R = correlational. EE = empirically equivalent. RAND = random. CONT = content. EFA = exploratory factor analysis.

structural paths and overall criterion variance explained. The SFA, R, EE, and RAND methods produce slightly weaker structural path coefficients and a lower explained criterion variance. It is important to note that these differences between the methods are greatest at larger sample sizes. At the smallest sample sizes there are few clear differences between most of the methods.

Table 6
Mean Standardized Structural Path Coefficients

	Model													
	TRIM		*SFA*		*R*		*EE*		*RAND*		*CONT*		*EFA*	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Model with *N* = 1,177														
JOBSAT	.477	.024	.000	.054	–.021	.039	–.008	.038	–.028	.037	–.004	.019	.013	.024
REASON	.199	.027	.201	.306	.209	.296	.313	.026	.346	.043	.386	.031	.145	.311
USOC	.337	.028	.167	.220	.172	.236	.290	.032	.290	.049	.270	.030	.090	.265
DECSAT	–.001	.015	.207	.305	.199	.292	.344	.027	.332	.023	.334	.027	.160	.373
TENURE	.004	.023	.018	.034	.028	.040	.036	.019	.036	.029	.034	.018	–.004	.067
Model with *N* = 300														
JOBSAT	.475	.083	.003	.062	–.029	.054	.020	.071	.018	.066	–.014	.037	.009	.041
REASON	.187	.068	.267	.235	.283	.222	.308	.138	.266	.228	.401	.052	.338	.104
USOC	.340	.056	.213	.165	.224	.168	.255	.077	.211	.203	.250	.070	.245	.066
DECSAT	.002	.036	.289	.245	.272	.241	.357	.070	.285	.223	.341	.071	.390	.042
TENURE	.004	.040	.018	.058	.028	.053	.041	.061	.016	.036	.034	.037	–.013	.035
Model with *N* = 100														
JOBSAT	.270	.132	.164	.122	.133	.143	.185	.180	.161	.130	.016	.062	.085	.173
REASON	.399	.257	.286	.162	.259	.244	.277	.252	.192	.257	.494	.226	.053	.343
USOC	.264	.210	.384	.122	.324	.216	.164	.460	.341	.258	.187	.225	.375	.832
DECSAT	.001	.086	.230	.135	.133	.230	.160	.227	.166	.196	.304	.132	.090	.248
TENURE	.087	.080	.158	.102	.106	.134	.159	.205	.104	.160	.014	.076	.117	.163

Note. JOBSAT = level of job satisfaction. REASON = reasons for joining the union. USOC = socialization experiences in the union. DECSAT = satisfaction with the union's decision-making process. TENURE = tenure. TRIM = trimmed. SFA = single factor. R = correlational. EE = empirically equivalent. RAND = random. CONT = content. EFA = exploratory factor analysis.

Discussion

The results of this study can be viewed as being complementary to the theory-based discussion of construct specificity provided by Bagozzi and Edwards (1998) and to the Hall et al. (1999) discussion of the influence of unmodeled secondary constructs on item parcels. Results of this study suggest that the methods used for creating composites produced models with varying degrees of fit. As Bagozzi and Edwards suggested, conceptual decisions with regard to appropriate measurement model specification can significantly influence the resulting overall model fit.

Several conclusions seem warranted from the preceding simulation results. First, composites of almost all types accomplish their goals. That is, model fit is substantially improved when composites are used as compared to treating all indicators individually. This effect is particularly pronounced at small sample sizes. In fact, with a more realistic sample size (i.e., 300), the use of composites resulted in sample size-to-estimated-paths ratios of approximately 7:1 to approximately 14:1. Because many studies have sample sizes that preclude using all indicators separately, the practice of composite formation not only appears warranted but it also appears justified in terms of yielding a set of well-fitted measurement models.

Second, the method of composite formation does have an impact on model fit. Specifically, the data suggest that the SFA, R, RAND, and EE techniques result in models that are clearly superior on model fit criteria to the other tested methods. Given the

similarity of fit produced by these methods, the RAND method may be the most appealing alternative in that no initial analyses need to be performed before creating composites. It is likely, however, that the success of the RAND method in this study was due to the unidimensionality of the items. If items in a scale measure multiple dimensions of the latent variable, it is unlikely that a random parceling of those items will produce parallel composites.

Another important consideration when evaluating the use of composites is to determine why some methods produce better fit than others do. The data suggest that model fit can be improved through the formation of composites. However, this improved overall model fit may be due to emphasizing the creation of strong measurement models at the expense of the structural relationships. The lambda values suggest that the composite methods produced better model fit by creating stronger measurement models relative to the TRIM model. Examining the average criterion variance explained by the various methods further supports this. Although the TRIM model had the lowest overall model fit and the lowest lambda values, the variance explained in the criterion by the structural relationships was noticeably larger than any of the composite methods. In short, although the underlying structural model may be incorrect, acceptable model fit may be produced through the creation of unduly strong measurement models.

Although perhaps the most conceptually pure, and despite the amount of criterion variance explained, the CONT method did not result in very good overall model fit at any of the tested sample sizes. There are several possible explanations for this in this study. First, for several scales, it was not readily apparent as to how the composites should be created. Specifically, the satisfaction-with-the-union scale could well have been defined as essentially unidimensional. Thus, the forced creation of composites may have resulted in psychometrically nonequivalent indicators. Second, if researchers were adept at the a priori identification of the underlying factor structure of a set of items, empirical factor analysis would be rendered moot. Thus, our categorizations may simply have not been the most appropriate.

Interestingly, the EFA method did not produce consistently good model fit. A closer examination of this technique using the current data suggests that one of the potential problems is that there is significant variation in the construction of composites. Within a given sample size, items were combined differently across the 10 simulations. That is, for one simulation, a particular latent variable may have three composite indicators, whereas in another simulation, only two composite indicators may have been formed. Although this is possible for all of the empirically driven methods, its effect was considerably more noticeable in the EFA technique. This phenomenon was most apparent at smaller sample sizes and is evidenced by the variation in the degrees of freedom (see Table 2).

A related issue concerns the extent to which other composite formation methods produced the exact same composite structures (i.e., same items within each composite). In particular, the similarity of results associated with the SFA, R, RAND, and EE techniques suggests the possibility that this was due to these methods producing similar, if not identical, composite structures. This is of particular concern for variables measured with relatively few items (e.g., four items for JOBSAT) due to the small number of possible combinations. Examination of composite structure, however, indicates that these methods did produce different composites. With respect to the JOBSAT variable, for example, when $N = 300$, the EE method produced the same two

composites (Set A) 9 times out of 10 and a different set of composites (Set B) on 1 simulation. Alternatively, the RAND method obtained composite Set A five times and composite Set B three times but also got a third composite (Set C) two times. Although some overlap would be expected based on the fact that there are only three combinations of two composites for this variable, these results suggest that different formation methods do produce dissimilar composites. For variables with more items (e.g., USOC and UCOMM), the degree of overlap was even less evident. In short, it does not appear that the similarity of results for the SFA, R, RAND, and EE methods was heavily influenced by the structure of the generated composites.

Possible Limitations and Future Research

One potential explanation of the observed results is that only a single data set was used and only 10 bootstrap samples were generated for each sample size. Compared with other simulation-based work, this *N* is dramatically less than would be preferable. An examination of the fit statistics, lambdas, and criterion variance accounted for across the 10 samples, however, suggests that there was relatively low within-method variability as compared to between-method variability.

It is also possible that the observed results were due to sample-specific factors associated with the set of data. An examination of the item characteristics (see Appendices A and B), however, does not suggest an obvious explanation for these results being idiosyncratic. In fact, these data appear similar (e.g., in terms of the dimensionality of the constructs, measurement characteristics of the items, and hypothesized theoretical relationships) to those often observed in the extant psychological and organizational literature. The use of observed versus simulated data is a key difference between this study and related work by Hall et al. (1999). In short, a limitation of this study is the inability to assess the extent to which the various composite formation methods allow for the accurate recovery of known parameters. There is a clear need for future research directed at integrating the results of this study with recent simulation-based work by Hall et al. and Marsh et al. (1998). Such research would likely provide important information to researchers with regard to decisions of if and how to form composites.

Another possible influential factor is related to the adoption of the Anderson and Gerbing (1988) approach to trimming measurement models before forming composites. Specifically, using this procedure may have resulted in a greater homogeneity of the retained items. This homogeneity, however, is not necessarily problematic. Rather, the domain-sampling model would suggest that the dimensionality of the composites be driven by the dimensionality of the initial item pool. If items in the original pool were homogeneous, then the composites should likewise be homogeneous. Because the Anderson and Gerbing procedure trims bad items, it is certainly possible that it could destroy the initial heterogeneity of an item pool. Although it probably depends on why an item is trimmed (i.e., low loading on primary construct *or* high cross-loadings with other constructs), it is not immediately clear as to how the technique would have influenced the results with respect to the domain-sampling model. Although this two-step method has been advocated by others (e.g., Medsker et al., 1994), a one-step model in which measurement and structural relationships are immediately tested simultaneously might produce different results. A potential avenue for future research would be to consider the possible effects of including items with

exceptionally poor loadings (or high cross-loadings) on the pattern of findings across the various composite formation methods.

Summary

The results of this study raise both practical and theoretical issues. On the practical side, if one is concerned with maximizing the ratio of sample size to estimated paths, some of these models appear to work more efficiently and effectively than others. The SFA, RAND, EE, and R methods result in more acceptable ratios of sample size to estimated paths than the total disaggregation model across all sample sizes as well as produce better fitting models. Although the CONT and EFA models may produce better ratios, they result in models with appreciably poorer fit. Although the SFA, RAND, EE, and R methods appear to be roughly equivalent to the extent that any given grouping of items in a scale will create psychometrically equivalent measures, the RAND method will be the easiest to implement. These results, however, should not be interpreted as suggesting that these four methods (i.e., SFA, RAND, EE, and R) should always be used to form composites in applications of SEM. Considering the results of this study in addition to other recent work in the area (e.g., Bagozzi & Edwards, 1998; Hall et al., 1999; Little et al., 1999), the decision to form composites should primarily be driven by the researcher's conceptualization of the focal construct. In short, although situational constraints may induce a researcher to use composite measures, issues related to construct specificity and purity should always be of paramount importance (Bagozzi & Edwards, 1998).

APPENDIX A
Descriptive Statistics for Items Used to Form Composites (*N* = 1,177)

Item	M	SD	*Skewness*	*Kurtosis*
JOBSAT1	3.23	1.16	–0.34	–0.99
JOBSAT2	4.05	0.93	–1.08	1.01
JOBSAT3	4.05	0.92	–1.12	1.19
JOBSAT4	3.75	1.02	–0.80	0.09
JOBSAT5	3.94	0.98	–1.01	0.63
JOBSAT6	3.32	1.06	–0.28	–0.54
REASON1	3.53	1.16	–0.77	–0.24
REASON2	4.12	0.93	–1.36	2.13
REASON3	3.69	0.97	–0.67	0.44
REASON4	3.84	1.01	–0.87	0.48
REASON5	3.36	1.26	–0.45	–0.80
REASON6	2.96	1.16	–0.13	–0.75
REASON7	4.02	0.95	–1.16	1.54
REASON8	4.02	1.01	–1.08	0.94
USOC1	2.90	1.06	0.05	–0.70
USOC2	3.56	0.99	–0.62	0.03
USOC3	3.95	0.92	–1.08	1.38
USOC4	3.57	1.03	–0.70	0.03
USOC5	3.63	1.00	–0.52	–0.11
USOC6	3.13	1.01	–0.27	–0.52
USOC7	3.31	1.02	–0.20	–0.58

(continued)

APPENDIX A Continued

Item	M	SD	*Skewness*	*Kurtosis*
USOC8	3.62	0.88	–0.60	0.62
USOC9	3.45	1.04	–0.42	–0.45
USOC10	2.13	0.98	0.75	0.06
USOC11	2.62	0.97	0.15	–0.41
USOC12	2.17	0.76	1.17	2.12
USOC13	3.44	1.00	–0.59	–0.27
DECSAT1	3.73	1.00	–0.95	0.52
DECSAT2	3.73	0.93	–0.88	0.66
DECSAT3	3.49	0.90	–0.52	0.20
DECSAT4	3.71	0.97	–0.85	0.52
DECSAT5	3.65	0.85	–0.45	0.47
DECSAT6	3.98	0.86	–0.86	0.95
TENURE	16.13	10.72	0.19	–1.20
UCOMM1	3.39	0.97	–0.40	–0.10
UCOMM2	3.54	0.91	–0.55	0.25
UCOMM3	3.57	0.94	–0.65	0.32
UCOMM4	3.78	0.87	–0.81	1.12
UCOMM5	3.16	0.95	0.01	–0.24
UCOMM6	3.82	0.78	–0.86	1.59
UCOMM7	3.38	1.03	–0.42	–0.50
UCOMM8	3.78	0.89	–0.77	0.67
UCOMM9	2.59	1.12	0.49	–0.50
UCOMM10	2.90	1.01	0.01	–0.39
UCOMM11	3.70	0.86	–0.68	0.63
UCOMM12	4.07	0.78	–0.85	1.43
UCOMM13	4.28	0.78	–1.32	2.77

Note. TENURE = tenure. JOBSAT = level of job satisfaction. REASON = reasons for joining the union. USOC = socialization experiences in the union. DECSAT = satisfaction with the union's decision-making process. UCOMM = level of commitment felt toward the union. All items were measured on 5-point scales with the exception of TENURE, which was expressed as the number of years individuals had been working as teachers. Italicized items are those that were dropped after performing the Anderson and Gerbing (1988) trimming procedure.

APPENDIX B
Factor Loadings From Exploratory Factor Analyses for *N* = 1,177

Scale	*Item*	*Factor 1*	*Factor 2*
JOBSAT (extracted one factor)			
	JOBSAT1	.53	
	JOBSAT2	.93	
	JOBSAT3	.86	
	JOBSAT6	.38	
REASON (extracted two factors)			
	REASON2	.72	
	REASON4	.63	.33
	REASON5		.59
	REASON6		.76
	REASON7	.51	.31
	REASON8	.50	

APPENDIX B Continued

Scale	*Item*	*Factor 1*	*Factor 2*
USOC (extracted one factor)			
	USOC1	.80	
	USOC2	.57	
	USOC4	.65	
	USOC5	.67	
	USOC6	.80	
	USOC9	.68	
	USOC10	.51	
	USOC11	.68	
	USOC13	.52	
DECSAT (extracted one factor)			
	DECSAT1	.69	
	DECSAT2	.86	
	DECSAT3	.79	
	DECSAT4	.76	
	DECSAT5	.69	
	DECSAT6	.75	
UCOMM (extracted two factors)			
	UCOMM1	.82	
	UCOMM3	.86	
	UCOMM4	.86	
	UCOMM6	.44	.45
	UCOMM8		.64
	UCOMM10	.60	.32
	UCOMM11		.82
	UCOMM12	.20	.66
	UCOMM13	.63	.26

Note. TENURE = tenure. JOBSAT = level of job satisfaction. REASON = reasons for joining the union. USOC = socialization experiences in the union. DECSAT = satisfaction with the union's decision-making process. UCOMM = level of commitment felt toward the union. Loadings less than .20 were not reported. For REASON and UCOMM, reported loadings are based on a rotated solution.

Notes

1. Data were initially generated and analyzed for sample sizes of 100, 150, 200, 250, 300, and 1,177. Because the results associated with all conditions were similar and in the interests of space, only data and results from three sample sizes (100, 300, and 1,177) are reported in this article. Results for all sample sizes are available on request from the first author.

2. Although the analysis of covariances is usually preferred in structural equation modeling, correlation matrices were used in this study for two reasons. First, in preliminary analyses, the use of covariances occasionally resulted in solutions that failed to converge. Furthermore, when they did converge, no differences in overall model fit results were noted in comparison to those produced from correlation matrices. Second, Cudek (1989) indicated that many models for latent variable regression are scale invariant and that analysis of correlations in some cases is acceptable. The models tested in this study are consistent with one of the cases described by Cudek in which the diagonal elements of the Φ matrix are fixed to unity, Ψ is unconstrained, and only fixed values of zero are introduced to Λ.

References

Agars, M., Unckless, A., & Tesluk, P. E. (1998, April). *A cross-level examination of factors influencing union commitment and participation*. Poster session presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45-87.

Barling, J., Fullagar, C., & Kelloway, E. K. (1992). Union loyalty and strike propensity. *The Journal of Social Psychology*, *132*, 581-590.

Barling, J., Wade, B., & Fullagar, C. (1990). Predicting employee commitment to company and union: Divergent models. *Journal of Occupational Psychology*, *63*, 49-61.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural equation modeling. *Sociological Methods and Research*, *16*, 78-117.

Brooke, P. P., Russell, D. W., & Price, J. L. (1988). Discriminant validation of measures of job satisfaction, job involvement, and organizational commitment. *Journal of Applied Psychology*, *73*, 139-145.

Cramer, D. (1996). Job satisfaction and organizational continuance commitment: A two-wave panel study. *Journal of Organizational Behavior*, *17*, 389-400.

Cudek, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317-327.

Fullagar, C., Clark, P., Gallagher, D., & Gordon, M. E. (1994). A model of the antecedents of union commitment: The role of socialization experiences and steward characteristics. *Journal of Organizational Behavior*, *15*, 517-533.

Gordon, M. E., Philpot, J. W., Burt, R. E., Thompson, C. A., & Spiller, W. E. (1980). Commitment to the union: Development of a measure and an examination of its correlates. *Journal of Applied Psychology*, *65*, 479-499.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *2*, 233-256.

Hammer, T. H., & Wazeter, D. L. (1993). Dimensions of local union effectiveness. *Industrial and Labor Relations Review*, *46*, 302-319.

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.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks, CA: Sage.

Kelloway, E. K., Catano, V. M., & Southwell, R. R. (1992). The construct validity of union commitment: Development and dimensionality of a shorter scale. *Journal of Occupational and Organizational Psychology*, *65*, 197-211.

Lent, R. W., Lopez, F. G., Brown, S. D., & Gore, P. A., Jr. (1996). Latent structure of the sources of mathematics self-efficacy. *Journal of Vocational Behavior*, *49*, 292-308.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4,* 192-211.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181-220.

Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology, 76,* 127-133.

Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal, 35,* 828-847.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management, 20,* 439-464.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management, 21,* 335-355.

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Williams, L. J., & Anderson, S. E. (1994). An alternative approach to method effects by using latent-variable models: Applications in organizational behavior research. *Journal of Applied Psychology, 79,* 323-331.

Ronald S. Landis is an assistant professor in the Department of Psychology at Tulane University. He received his Ph.D. from Michigan State University. His current research interests are in the areas of applied research methods and personnel selection.

Daniel J. Beal is currently a doctoral candidate in the Department of Psychology at Tulane University. His current research interests are in the areas of aggression and stereotyping.

Paul E. Tesluk is an assistant professor in the Department of Management and Organization at the University of Maryland. He received his Ph.D. from Pennsylvania State University. His current research interests are in the areas of design and implementation of high-involvement workplace systems and work team performance.