To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits

Todd D. Little , William A. Cunningham , Golan Shahar & Keith F. Widaman

To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits

Todd D. Little, William A. Cunningham, and Golan Shahar
Department of Psychology
Yale University

Keith F. Widaman
Department of Psychology
University of California at Davis

We examine the controversial practice of using parcels of items as manifest variables in structural equation modeling (SEM) procedures. After detailing arguments pro and con, we conclude that the unconsidered use of parcels is never warranted, while, at the same time, the considered use of parcels cannot be dismissed out of hand. In large part, the decision to parcel or not depends on one's philosophical stance regarding scientific inquiry (e.g., empiricist vs. pragmatist) and the substantive goal of a study (e.g., to understand the structure of a set of items or to examine the nature of a set of constructs). Prior to creating parcels, however, we recommend strongly that investigators acquire a thorough understanding of the nature and dimensionality of the items to be parceled. With this knowledge in hand, various techniques for creating parcels can be utilized to minimize potential pitfalls and to optimize the measurement structure of constructs in SEM procedures. A number of parceling techniques are described, noting their strengths and weaknesses.

Using parcels as indicators of constructs in structural equation models (SEMs) has been and remains a controversial practice. Historically, debates on the utility and efficacy of parcels date back over 40 years (e.g., Cattell, 1956; Cattell & Burdsal, 1975), and the debates have continued in contemporary SEM circles (e.g., Bandalos & Finney, 2001; Kishton & Widaman, 1994; Marsh, Hau, Balla, & Grayson, 1998; SEMNET, 2001). The goals of this article are to (a) examine the pros and cons of the practice of parceling; (b) detail the uses, and attendant advan-

tages and disadvantages, of various parceling techniques; and (c) discuss the conditions under which parceling is warranted.

Parceling is a measurement practice that is used most commonly in multivariate approaches to psychometrics, particularly for use with latent-variable analysis techniques (e.g., Exploratory Factor Analysis, SEM). A parcel can be defined as an aggregate-level indicator comprised of the sum (or average) of two or more items, responses, or behaviors. To some, aggregating items to manufacture indicators of constructs is viewed as a dubious practice at best and cheating at its worst. Moreover, the practice of parceling contributes to the oft-whispered reputation of SEM as yielding a "smoke-and-mirrors" distortion of reality. For advocates of parceling, on the other hand, the practice is viewed as one that puts a fine sheen on an otherwise cloudy and therefore difficult to discern picture of reality. In this sense, the use of parcels in SEM is not seen as invoking smoke and mirrors, but rather as providing a carefully polished mirror of reality that really "smokes."

In our efforts to tackle this controversy, we have organized our thinking along three general lines. First, we assemble the arguments, both pro and con, for using parceling techniques and carefully detail the psychometric principles involved. Second, we describe a number of parceling techniques that can be used in specific situations. Third, we discuss various strategic considerations involved in the decision to parcel or not to parcel items for use in SEMs.

ARGUMENTS PRO AND CON

Common Theoretical Concerns

Most of the arguments both pro and con have focused on the differential analytic behavior and psychometric characteristics of items and parcels. However, philosophical arguments can also be levied. On one hand, the decision to parcel or not to parcel can be rendered moot if a researcher's philosophical position is on the con side. From an empiricist–conservative philosophy of science perspective, parceling is akin to "cheating" because modeled data should be as close to the response of the individual as possible in order to avoid the potential imposition, or arbitrary manufacturing, of a false structure. Any potential source of subjective bias on the part of the data analyst is to be avoided at all costs, a simple a priori principle with which most would agree. In this sense, allowing the researcher to create parcels from items fundamentally undermines the objective empirical purpose of the techniques that have been developed to model multivariate data.

From a more pragmatic-liberal philosophical perspective, parcels have potential merits as the lowest level of data to be modeled. Given that measurement is a strict, rule-bound system that is defined, followed, and reported by the investigator, the level of aggregation used to represent the measurement process is a matter of

choice and justification on the part of the investigator. With a compelling justification, using parcels would not be seen as a transgression against truth because the operational aspects of an investigation are, fundamentally, a public process. Editors, reviewers, and, ultimately, the field must approve the methods chosen by the investigator in order for the study's results to be accepted as veridical.

The gray areas between these two extreme positions have provided the fodder of debate: What constitutes a compelling justification for using parcels, and under what circumstances are arguments supporting the use of parcels tenable?

A common point of concern underlying many of these arguments is the nature of constructs and measurements in the behavioral and social sciences. Constructs can vary in their fundamental composition along two orthogonal dimensions. The first dimension is the continuum from unidimensionality to multidimensionality, or homogeneous to heterogeneous, respectively. Facility with adding basic digits, for example, represents a unidimensional construct, whereas intelligence at the level of "g" represents a multidimensional or heterogeneous construct. The second dimension along which constructs vary is the continuum from high to low explicitness. Descriptions of certain constructs are highly explicit, with a clear demarcation of the boundaries of the construct. For example, action-control beliefs are highly circumscribed both in terms of their theoretical definition and their empirical operationalization (Skinner, 1996). Other constructs are defined in a much less restricted manner, implicitly subsuming additional content "of the same type." For example, self-efficacy is quite variable across operational measurements of the construct (Multon, Brown, & Lent, 1991).

Measurements also vary in their fundamental nature along two orthogonal dimensions. The first dimension is the same as for constructs, a continuum from unidimensional (or homogeneous) to multidimensional (or heterogeneous). If a construct is unidimensional, then the measurements of that construct will contain unidimensional item content, as this is the only choice open. However, if a construct is multidimensional or heterogeneous in nature, various choices are open or viable. The investigator may choose to assess the relatively unidimensional or homogeneous core of the construct or may opt to obtain representative assessments of all facets of the behavioral domains covered by the construct. For example, in the NEO–personality inventory assessing the Big 5 personality factors, Costa and McCrae (1992) identified six facets for each of the Big 5 dimensions. The facets comprising the construct extroversion are gregariousness, assertiveness, activity, excitement seeking, positive emotions, and warmth (Costa & McCrae, 1992). One could measure the core facet (i.e., gregariousness) or select broadly across all facets when representing extroversion as a construct (Little, Lindenberger, & Nesselroade, 1999).

The second dimension along which measurements vary is a continuum that may be characterized as "clean to dirty." This continuum captures problems associated with the relative presence of unwanted sources of variance, such as method con-

tamination, acquiescence response bias, social desirability, and experimental effects such as fatigue and boredom. This second dimension reflects the various threats to validity associated with the design and execution of measurement operations. Here, a clean measure of a construct is one that is relatively uncontaminated and unconfounded by unwanted influences, whereas a dirty construct would be rife with unwanted sources of systematic error variance.

Given these dimensions along which constructs and measurements vary, the preceding philosophical positions appear to arise from differing interpretations of the unidimensional versus multidimensional dimension on which both constructs and their measures vary. The empiricist–conservative position appears rooted in the stance that all sources of variance in each item must be represented in any multivariate statistical models involving a given scale. Failing to represent one or more sources of variance, however minor, may lead to bias in estimates of other key parameters throughout the model. In contrast, the pragmatic–liberal position holds that representing each and every source of variance in each item, particularly on an a priori basis, is impossible. Under this position, researchers cannot know every single source of variance in every single item; one can only hope that one's models will represent the important common sources of variance across samples of items. When minor influences, which are substantively trivial yet empirically significant, cannot be predicted on an a priori basis, they will be difficult or impossible to distinguish from chance findings. Rather than engage in data snooping that is of questionable value, the pragmatic–liberal position would contend that researchers should concentrate on building replicable models based on solid and meaningful indicators of core constructs that will replicate across samples and studies.

The Empirical Pros of Parcels

Turning from theory to empirical issues, at least two classes of argument in favor of parcels have been offered (Bandalos & Finney, 2001). The first class, or category, focuses on the differing psychometric characteristics of items and parcels. The second category of argument has focused on the factor-solution and model-fit advantages accruing to models based on parcels. As we will see, each class of argument for parcels is, for the most part, an argument against items.

Regarding the first category of concern, numerous researchers have highlighted the psychometric merits of parcels relative to items. Compared with aggregate-level data, item-level data contain one or more of the following disadvantages: lower reliability, lower communality, a smaller ratio of common-to-unique factor variance, and a greater likelihood of distributional violations. Items also have fewer, larger, and less equal intervals between scale points than do parcels (see Bagozzi & Heatherton, 1994; Kishton & Widaman, 1994; McCallum, Widaman, Zhang, & Hong, 1999; cf. Hau & Marsh, 2001). As we discuss below, these concerns are grounded in basic psychometric theory.

A second category of concern focuses on the number of parameters required to model items versus parcels and on the overall fit of structural models. Advocates argue that, because fewer parameters are needed to define a construct when parcels are used, parcels are preferred, particularly when sample sizes are relatively small (e.g., Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994). Marsh and Hocevar (1988) argued that the item:subject ratio be explicitly considered because lower ratios may lead to instability of the factor solution, particularly if the psychometric properties of the items are poor. A similar argument focuses on over-all model fit. Simply stated, the various indexes of model fit are expected to be more acceptable when parcels, rather than items, are modeled because of the psychometric and estimation advantages of parcels. Compared with item-level data, models based on parceled data (a) are more parsimonious (i.e., have fewer estimated parameters both locally in defining a construct and globally in representing an entire model), (b) have fewer chances for residuals to be correlated or dual loadings to emerge (both because fewer indicators are used and because unique variances are smaller), and (c) lead to reductions in various sources of sampling error (MacCallum et al., 1999).

Psychometric considerations.   As can be seen, the different arguments in favor of parcels over items are not identical because differing psychometric principles underlie each of them. In order to illustrate the merits of these views, we present a brief overview of classical test theory and the principles of aggregation.

In their now classic study, Rushton, Brainerd, and Pressley (1983) discussed the pitfalls associated with the use of a single item to represent a psychological construct. They argued against items, relative to aggregate scores, on two bases. First, they pointed out that individual items are unlikely to be as representative of the construct that a researcher wants to measure as would an aggregate score (i.e., a selection rationale). Second, they argued that individual item scores are statistically less reliable than aggregate scores (i.e., a psychometric rationale). Using published data, they demonstrated that, in each of several instances, theoretically expected relations emerged between conceptually similar constructs when all variables were included as aggregate scores; however, this was not the case when individual items were used. Rushton et al. concluded that the disregard for the principles of data aggregation had led to improper inferences and hampered progress in the field. The logic of their arguments regarding the problems with items remains principally the same logic used by those who advocate for parcels.

As an illustration using classical test theory, the domain-sampling model discussed by Little et al. (1999) provides a framework by which the principles of aggregation can be understood. The three fundamental assumptions of the domain-sampling model are that (a) a construct exists, (b) an infinite number of indicators can be selected to measure the construct, and (c) each indicator has some degree of association with the construct's true centroid. Although we rely on

the domain-sampling model for this discussion, each proposition and conclusion can be supported via alternative metaphors and methods (see Little et al., 1999). Any given indicator of a construct can thus be represented as

$$X_i = T_i + S_i + e_i$$

where X_i represents the score of an individual for a particular item, T_i represents the target construct component, S_i represents the idiosyncratic or specific component, and e_i represents random error. In other words, a given item is assumed to contain various sources of variance: a "true" core aspect (i.e., the part of an item that assesses the construct we desire to measure), a "specific" component (i.e., the idiosyncratic, but reliable component of an item that is unrelated to the construct), and a random error component (i.e., the theoretically meaningless "junk" or noise). Figure 1, which is based on Little et al., illustrates the relations among the three conceptual sources of variance that comprise an item in relation to the construct that the item indicates. From this conceptualization, a number of basic propositions emerge. For example, because e_i contributes to the total variance of an item, any statistical indexes that rely on estimates of shared variance, such as correlations and regression coefficients, will be underestimated if analyses are based on items (Nunnally, 1978). Moreover, both the random error and the specific components of an item reduce its communality (Gorsuch, 1983).

Such problems are remedied through aggregation. If all the infinite indicators of the construct were measured, the random errors, e_i, and the specific components,
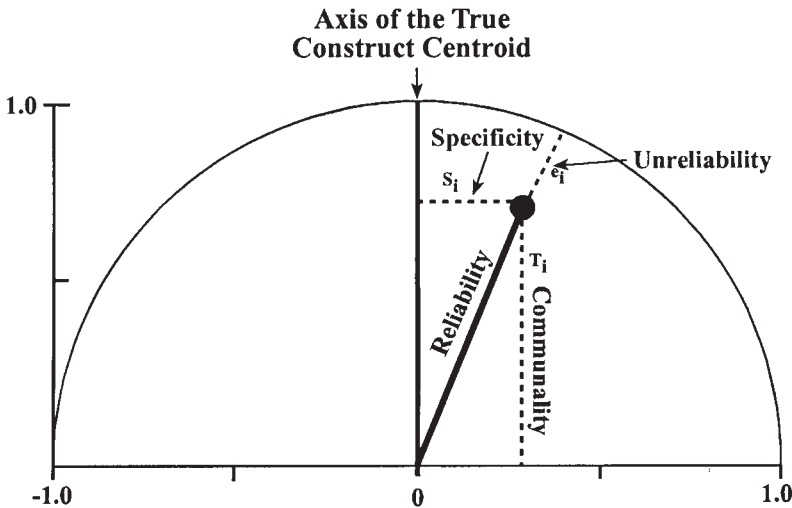

FIGURE 1   Geometric representation of the sources of variance of an indicator.

S_i, would be drastically reduced because e_i and S_i are defined as uncorrelated sources of variance within an item and, across all items in a domain, the e_i and S_i components are also assumed to be uncorrelated. As a result, an aggregate indicator that is the sum of an infinite number of items would consist almost completely of the geometrically compounding variance of the true component of the construct core, T_i. In practice, however, because only a subset of items is selected to represent a construct, some S_i variance may overlap with other indicators of similar composition (i.e., items that are selected from the same general quadrant of a domain). For example, a subset of items may share a method component, a response bias such as social desirability, or simply excessive item overlap, leading to what Cattell (1961) referred to as a "bloated specific" and what we have described as a "dirty" measure of a construct.

Unlike random error (e_i), the specific factor of an item (S_i) is reliable and conceptually reflects the presence of another construct, however narrow and irrelevant to the core definition of the latent variable of primary interest. Although we typically assume that each S_i has a mean of zero, is normally distributed, and is uncorrelated with other specific factors, the S_i components of items will not necessarily cancel out if the item-selection process is at all biased (Little et al., 1999). In other words, systematic variance unrelated to the latent construct of primary interest can create problems when an item correlates with one or more other items that share the same specific component (e.g., a method factor, a social desirability factor). Like the primary construct of interest, specific factors S_i of items can be multidimensional, such that the S_i's across a number of items share only a facet of their total unique sources of variance. Furthermore, the number of sources of variance that can contaminate a particular measured variable is practically limitless.

The law of large numbers, typically discussed in terms of the efficiency of parameter estimates (e.g., of a population mean), also holds for indicators of constructs. Within a person, a person's true score is more confidently represented to the extent that a larger number of measurements of the construct are used. Not only does the law of large numbers suggest that more items are better than fewer items in estimating a construct centroid, but it also suggests other ways in which aggregate scores are preferable to item scores. For example, as the number of items increases, nonnormal distributions become more normally distributed. As such, item distributions, which may have problems with skewness and kurtosis (thus violating assumptions of statistical inference), become more normally distributed when aggregated into scale scores or parcels. Similarly, scale intervals increase in number and effectively become both smaller and more equal with regard to the distances between points as more items are aggregated. For instance, the aggregation of two items, each measured on 4-point scales, yields a new, parceled indicator that has seven scale points; because of the normalizing tendency of aggregation, the intervals would become, of necessity, smaller and more continuous in nature. Each change of 1 scale point on the parcel-level scale encompasses a smaller proportion

of the cumulative distribution of scores than a change of 1 scale point on the item-level scale.

A related psychometric consideration regarding measurement is the distinction between the bias of an estimate and its efficiency. Bias refers to the "on average" behavior of items versus parcels to reveal the true centroid of a construct. Efficiency refers to the variability in this on average behavior. Any study reflects a single instantiation of a selection of indicators to represent a construct. Fundamentally, a given study contains only one selection of indicators out of the infinite number of sets of indicators possible. From a bias perspective, a given selection of indicators will, on average, reflect the construct with a given degree of bias. However, from an efficiency perspective, the one instantiation has a greater likelihood of missing the target if efficiency is low and parameter estimates are therefore more variable. The simulation and taxonomic model presented by Little et al. (1999) demonstrated that indicators selected from less diverse domains are more efficient than those selected from more diverse domains. As illustrated in Figure 2, the diversity of parceled indicators can be considerably less than the diversity of item-level indicators, and the communality of the parceled indicators can be considerably greater. Although analyses based on both sets of indicators would have the same implications for bias (i.e., on average, both would lead to equally accurate construct representation), the analyses based on parcels would be more efficient than the item-level analyses. The efficiency of parceled indicators implies that, if the selections are off base, they will not be as wildly or variably off base as would item-based indicators.

This effect can be seen in Figure 3. Panel A of Figure 3 displays a hypothetical scattering of selected items around the centroid of a construct. Parceling items into groups of three, for example, would result in a reduced diversity of indicators surrounding the centroid as is seen in Panels B through F of Figure 3. When the variances of the items are equivalent, the location of a parcel in construct space is the geometric center of the area formed from the chosen items (for two items, the parcel would be located at the geometric midpoint of the line between any two items that are paired). Figure 3 indicates that even a random process of selecting pairs, triads, quadrads, or even larger number of items for parcels leads to a tighter, less diverse, and, therefore, more efficient construct space.

A general conclusion that can be drawn from the foregoing psychometric considerations is that the use of additional items yields a more encompassing and inclusive representation of a construct. However, this increase in number of items can create problems in SEM. Practically speaking, specifying a latent variable with a large number of indicators poses numerous problems, as discussed in the next section.

Model-level considerations.   A first problem related to model-level considerations is related to Type I error. Due to the fact that about 1 in 20 correlations

Lawrence Erlbaum Associates, Inc. does not have electronic rights to Figure 2. Please see the print version.

FIGURE 2   Geometric representation of the psychometric differences between selecting items versus parcels for representing constructs. From "On Selecting Indicators for Multivariate Measurement and Modeling With Latent Variables," by T. D. Little, U. Lindenberger, and J. R. Nesselroade, 1999, *Psychological Methods, 4,* p. 197. Copyright 1999 by the American Psychological Association. Adapted with permission.
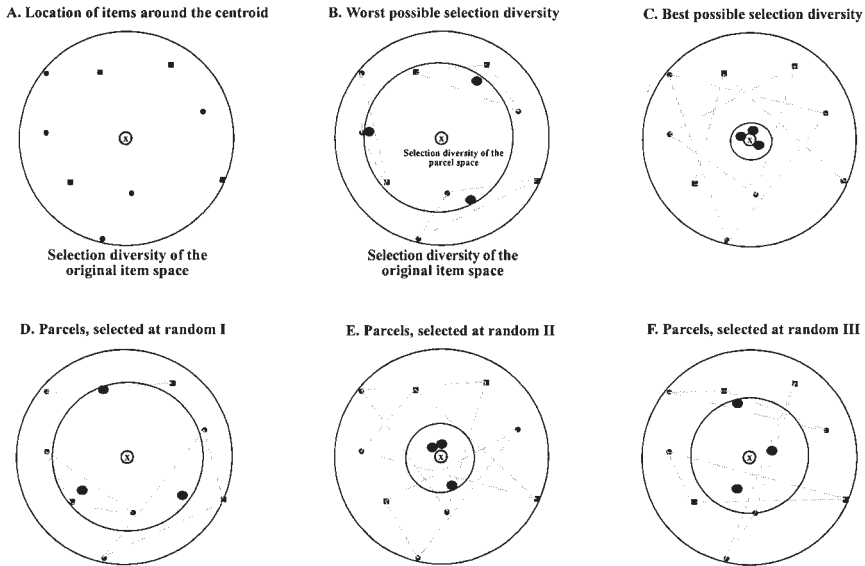
FIGURE 3   Geometric representation of the changes in selection diversity as a function of which items are parceled.

would be significantly related by chance alone (at a .05 Type I error rate), a 10-variable correlation matrix, with $10(9)/2 = 45$ unique correlations, would contain about two spuriously significant correlations even if all correlations in the matrix were zero in the population. By chance alone, and assuming no bias in indicator selection, a model with three constructs each measured with 10 variables would yield about 22 spurious correlations, whereas a structural model with three constructs, each measured with three parcels of items, would yield only about two spurious correlations. The nature of the spurious correlations could manifest as unreplicable relations among residuals or dual-factor loadings. In either instance, failing to estimate these significant, yet spurious relations would lead to poorer model fit, whereas estimating them would lead to false interpretations.

We should note that this problem of spurious levels of covariation among variables holds even if the population correlation between variables is not zero. The mathematical developments underlying factor analysis and structural modeling require assumptions that the specific and error components (S_i and e_i, respectively) are uncorrelated with one another and with the true target construct component, T_i. However, at best, these assumptions are satisfied only in the population and will not hold in any finite sample from the population (MacCallum et al., 1999). Thus, even in fairly large samples for the behavioral sciences (e.g., 400 participants), some nonzero covariation among the true, specific, and error components would

be expected to occur. These nonzero covariances contaminate the correlation between two items; if sufficiently large, they would lead to the necessity to specify a correlation between the unique factors. Moreover, given the evanescent nature of these added covariances, they would not be expected to recur in a new sample, rendering the specification of parameter estimates for these influences to be mere data snooping or model fitting.

A second problem related to selecting and using more items as indicators of constructs is the attendant likelihood that subsets of items will share specific sources of variance. These specific sources of variance themselves represent latent constructs no matter how narrow and trivial their nature. Such constructs are unlikely to be hypothesized by the researcher in an a priori manner and would lead to systematic sources of common variance that are not represented in the initial, a priori model specification. This problem can be thought of as resulting in misspecified a priori models. For advocates of parcels, the source of the misspecification is seen as an unwanted contamination resulting from the use of large numbers of items (however, see the con perspective, below). As such, parceling the items into fewer indicators would likely eliminate or at least reduce the unwanted source or sources and would lead to better initial model fit than if the items were used as indicators of constructs.

A third problem with more indicators per construct is related to the stability of solutions; because of the typically poor psychometric characteristics of items, item-based solutions are often unstable and take more iterations to converge, yielding relatively large standard errors of the measurement-level parameters (and generally poor model fit). In such situations, even small changes to a model can lead to noticeable changes in the magnitudes of parameters (although these changes may not affect the significance of the parameters). As a result, the generalizability of the parameters is compromised. Again, models based on parcels would not suffer from such noticeable effects.

Marsh et al. (1998) recently conducted a Monte Carlo study that is relevant to the "more-is-better" issue. In Study 3 of their simulation, they compared confirmatory factor analyses based on either two, three, four, six, or 12 items per latent construct and two, three, four, or six parcels created from all 12 items per latent construct. Although Marsh et al. acknowledged the "good" performance of a three-parcel solution, they argued that the 12-item solution was modestly better than the parcel solution and concluded that the more indicators used in a confirmatory factor model the better is the assessment of a latent construct. A proponent of parceling, however, may counter that the simulation was not a fair test. In their simulations, Marsh et al. used extremely well conditioned data. For example, the lowest magnitude of item loadings was around .6 for all items on all factors. The simulations also did not involve adding systematic error into the data (i.e., the S_i component described above thereby removing any chance of spurious effects). Under conditions in which one is modeling such clean and well-conditioned items in a

confirmatory factor context, the item versus parcel distinction carries little importance. However, given the pro arguments outlined earlier, under conditions in which one is modeling less-than-optimal items, the problems and pitfalls associated with typical item-level data would still be relevant.

A final issue related to the "more items" question has to do with the optimal number of indicators needed to identify and represent a latent construct. Even though latent constructs can be represented using one or two indicators (e.g., Holahan & Moos, 1994), such a practice is generally viewed as suboptimal because it reflects a locally underidentified latent variable.[1] On the other hand, three indicators of a construct lead to a just-identified latent variable whereas four or more indicators lead to an overidentified latent variable. A just-identified latent variable is arguably better than an overidentified one. First, a just-identified construct has only one unique solution that optimally captures the relations among the items, no matter what other constructs are considered or included in a model. Overidentified models can have more than one optimal solution, depending on the nature of the other constructs that are represented. By way of metaphor, on uneven terrain, a four-legged chair will wobble (i.e., more than one position is possible), but a three-legged stool will stand fast (i.e., it has only one unique sitting position). Depending on the number of items that have been measured to represent a construct, parcels can be used to effectively reduce the number of indicators to an optimal, just-identified level.

The Empirical Cons of Parcels

Opponents of the practice of parceling have countered the many pro arguments by focusing on two main areas. The first area of concern surrounds the dimensionality of a construct and the potential pitfalls of a misspecified factor model. The second area focuses on the meaning of parameter estimates, particularly if established norms are masked or distorted by the construction of parcels. Although the amount

---

[1]With respect to measurement error, a single-indicator latent variable is essentially equivalent to a manifest variable. In this case, the error of measurement is either fixed at zero or fixed at a non-zero estimate of unreliability; additionally, a second corresponding parameter would also need to be fixed because of issues of identification. Two indicators of a construct also pose a problem because it too is an underidentified situation. With two indicators, five parameters are needed to represent the construct (two error terms, two loadings, and a latent variance term); however, only three observed statistics are available to identify these parameters. Given that for every latent construct one parameter is automatically fixed to a nonzero value in order to set the scale of estimation, this still leaves four parameters and three observed statistics (two variance estimates and one covariance estimate). Although some would argue that the underidentified parameter can be estimated using other relationships in the observed variance–covariance matrix, others have argued that an additional, meaningful constraint should be placed on the parameters to be estimated. Based on their simulation study, Little et al. (1999) strongly recommended placing an equality constraint on the two loadings associated with the construct because this would locate the construct at the true intersection of the two selected indicators.

of argumentation for the pro side far outweighs the con side, the importance of the con arguments is not disproportionately weaker.

Multidimensionality and model misspecification.   When constructs are not unidimensional, and when it is unclear what dimensions may underlie a construct, espousing item parceling may be particularly problematic. In fact, Bagozzi (Bagozzi & Heatherton, 1994), Bandalos (Bandalos & Finney, 2001), and others have stated that only under conditions of unidimensionality should parceling be considered.

One line of reason supporting this caution is that parcels drawn from items assessing a multidimensional construct are themselves likely to be multidimensional in composition. Using multidimensional parcels can distort measurement models because they may provide biased loading estimates and make it difficult to interpret the nature of the variance of a latent construct. Due to the dimensional composition of the parcels and the resulting construct, the structural relations among latent variables would be very difficult to interpret. The difficulty in interpretation arises when the subdimensions of a construct are not highly correlated with each other. Using a domain-representative parceling technique (Kishton & Widaman, 1994), parcels can be created that combine relatively independent sources of variance into a latent variable. Any associations of such latent variables with others in a model would be susceptible to alternative explanations (i.e., the researcher would be unsure as to which dimension or source of variance produced the structural effect). The bottom line for this viewpoint is that, when a latent variable is defined with multidimensional parcels, one can never be completely sure as to what the latent construct "really" is.

Problems associated with multidimensionality are not always readily apparent. Generally, we think of multidimensionality resulting from an item that is affected by two or more substantive constructs; such an item, therefore, is a measure of the constructs that affect it. Problems involving multidimensionality for structural equation modeling occur when fewer dimensions than exist in the data are specified (i.e., the model is misspecified with too few constructs). For example, a specific factor, S_i, may be thought of as systematic error, as opposed to random error (e_i), because it represents reliable variance that is unrelated to the latent construct of interest. However, this systematic error can also be considered as defining a dimension, particularly when it is shared across two or more items in a data set. Problems of unmodeled multidimensionality can arise from systematic error being shared across various items that are spread across the various parcels that define a given latent variable.

Latent variables that are modeled with parcels that share systematic error become defined, in part, by that systematic error. Measures that include variance associated with a core construct of interest and shared systematic error are thereby confounded, resulting in a confounded latent construct. In interpreting relations

with this construct, relations among the confounded latent variable and other constructs in one's model will reflect the influence of the core construct we intended to measure, the systematic error, or a combination of the two, leading to problematic interpretation. With confounded latent variables, we should expect latent relations to be underestimated when the systematic error is not present in other latent variables, but we should expect the correlations to be overestimated when the systematic error is shared in other variables. When such sources of systematic error are spread across parcels, this error becomes part of the latent construct.

As we have already noted, parcel-based models attempt to cancel out random and systematic error by aggregating across these errors, thereby improving model fit (Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994). The advantages and appropriateness of this approach to increasing model fit is controversial (e.g., Hall, Snell, & Foust, 1999). Bandalos and Finney (2001) argued that parceling would improve model fit for all models, correctly specified or not. As such, parceling may reduce our ability to identify misspecified models and may increase our Type II error rate (failing to reject a model that should be rejected). For example, Bandalos (1997) found that parceling could mask double loadings of items.

Generally speaking, parceling can hide many forms of misspecification, at least misspecification that would be found if analyses were performed on item-level data. As mentioned, the pro-parcel argument suggests that the hidden sources of error may effectively be removed from the to-be-analyzed data. A model fitted to parcel-level data would, therefore, no longer be misspecified. The con-parcel argument would counter that the use of parceling to remove the unwanted errors fundamentally changes the reality of the data such that the fitted model is a misrepresentation. To avoid such situations, some have argued that a more defensible strategy would be to model data at the item level in order to examine possible sources of misspecification (e.g., Bandalos & Finney, 2001).

Established norms.   Even in cases when parceling is a defensible statistical procedure, employing parcels may still be ill advised on applied grounds. Throughout the behavioral and social sciences, many scales have established norms based on their means, standard deviations, clinical cutoff scores, and so on. In clinical research, for example, several scales that tap the same theoretical construct are commonly employed (e.g., depression; see Tanaka & Huba, 1987). These scales may be used as indicators of a latent construct. When many scales are administered, one may be tempted to parcel them into fewer indicators of the target construct. Although this procedure may have statistical advantages (especially when the unidimensionality of this construct was previously established), it may run the risk of losing important applied information that is contained in each scale. For example, scale-level information may help assess the relative severity of the sample studied because both the intercepts and slopes of the scales are readily interpretable when represented in their original untransformed metrics.

The "established norms" argument suggests that the unstandardized parameters are meaningful in terms of threshold parameters or symptomatology and that parcels would create an arbitrary metric that would no longer carry this information. In addition, interpretable unique effects of scales are possible when established scales are used as indicators of a latent construct. In other words, the established norms argument suggests that the unique effects of established measures may also have important theoretical or clinical relevance that would be missed if parcels were used. This line of reasoning is less applicable to the item versus parcel debates and is more relevant to debates about scales versus parcels.

All of the arguments, both pro and con, have merits. However, before we turn to a detailed evaluation of the relative merits, we first provide a brief presentation of the various techniques that can be used to create parcels.

TECHNIQUES FOR BUILDING PARCELS

The various techniques that are available to build parcels generally share a common prerequisite: The dimensionality of the items to be parceled must be determined prior to parceling. As Bandalos and Finney (2001) pointed out, few studies that use parceling techniques report the dimensionality of the items used to represent a construct. Omitting such information, particularly for constructs whose dimensionality has not been demonstrated in the literature, is an unwarranted practice. If the dimensionality of a set of items is not known, the items could be prescreened using an exploratory factor analysis algorithm that uses an iterative estimator with an oblique rotation (i.e., the analogous algorithm of the SEM measurement model). However, we suggest verifying the presumed dimensionality of a set of items, particularly when the items are used across diverse and new populations. Once the dimensionality of a set of items is determined, then one or another of several techniques for parceling items can be applied. We turn first to techniques for unidimensional item sets.

Random Assignment

Following from a domain sampling rationale, one simple method for constructing parcels is to assign each item, randomly and without replacement, to one of the parcel groupings. Depending on the number of items to be assigned, two, three, or possibly four parcels, or groupings of items, could be created. Random assignment of items to parcels should, on average, lead to parcels that contain roughly equal common factor variance. We mention that items should generally stem from a common pool, such as questionnaire items responded to on a common scale. If the items evince unequal variances because the scales, or metrics, differ across items, the resulting parcel would be biased in favor of the items with the larger variances.

Of course, standardizing items to a common variance metric would alleviate such problems.

Item-to-Construct Balance

In constructing parcels for use in a larger SEM model, one goal is to derive parcels that are equally balanced in terms of their difficulty and discrimination (intercept and slope). If the mean levels of the indicators were of little or no concern, a simple examination of the item-to-construct relations would allow one to build balanced parcels. Specifically, one would specify a single-construct model that includes all items associated with the construct. Using the loadings as a guide, one would start by using the three items with the highest loadings to anchor the three parcels. The three items with the next highest item-to-construct loadings would be added to the anchors in an inverted order. The highest loaded item from among the anchor items would be matched with the lowest loaded item from among the second selections. If more items were available, the basic procedure would continue by placing lower loaded items with higher loaded parcels. Under some conditions, parcels may have differential numbers of items in order to achieve a reasonable balance.

Under conditions in which the intercept information is also important, this procedure can be extended to include the intercepts by specifying a single-construct model as mentioned previously, but request that the means, or intercepts, be estimated. In this case, one has to consider the relative balance between the discrimination parameter of the item (i.e., its loading) and its difficulty parameter (i.e., its intercept) in constructing balanced parcels.

As with any technique for constructing parcels, the item-to-construct relations should be verified in each potential subgroup that may be relevant (e.g., by gender, by age, by ethnicity, etc.). Some argue that item-to-construct relations should be verified anew in every new sample (Little, 1997). However, the most important advantages accruing to the use of structural equation modeling are present only if measures and analyses are constructed on a priori bases. As a result, once the item-to-construct relations have been verified several times in representative samples drawn from a given population, researchers can capitalize on this work to parcel items in a manner informed by the previous research when drawing participants from that population. Having well-established parcel indicators of constructs would enable the researcher to evade any arguments that the form of parceling was biased by too much data snooping in this sample.

A Priori Questionnaire Construction

Recent questionnaire designs presented by Little and colleagues have contained a priori guidelines for constructing parcels for use in SEM and related procedures (e.g., Little, Oettingen, & Baltes, 1995; Little & Wanner, 1997). For example, the

Control, Agency, and Means–Ends Interview (CAMI; Little et al., 1995) contains six items for each agency construct. Three of these items are worded in the negative direction (e.g., "I'm just not very smart"), and the remaining three are worded in the positive direction (e.g., "I can try hard"). In their instructions, Little et al. recommended that when creating parcels a positively worded item be coupled with a negatively worded item that has been reverse coded (thereby rendering the high scores as indicating high agency). The rationale for this recommendation is to reduce the negativity versus positivity bias, or acquiescence bias, relative to the underlying construct information regarding agency.

One important caveat about using a priori recommendations must be addressed. Specifically, item responses should be screened to ensure that they conform to the expected pattern. The creators of the CAMI instrument found a condition under which the expected structure of a construct did not conform. Specifically, in a sample of Japanese children, the three negatively worded luck items correlated positively with the three positively worded luck items (typically these correlations are negative prior to inversion; Karasawa, Little, Miyashita, Mashima, & Azuma, 1997). This quite unexpected finding prompted further inquiry and lead to the conclusion that luck may not be a unidimensional concept in Japan. Instead, it appears to have two faces: good luck and bad luck. A lucky person in this sociocultural context would be one who wins the lottery only to be hit by a bus. Thus, investigators must remain vigilant to variables that can potentially moderate the unidimensionality of a set of items.

Approaches to Multidimensionality

Kishton and Widaman (1994) described two methods for dealing with multidimensional item sets. To illustrate, imagine a nine-item scale comprising three facets (A, B, and C), each measured by three items (e.g., A_1 through A_3). The first approach, which Kishton and Widaman defined as the internal-consistency approach, creates three parcels that use the facets as the grouping criteria. The first parcel would reflect Facet A and would be the sum or average of A_1, A_2, and A_3. The second parcel would reflect Facet B, and the third would reflect Facet C. This approach would result in a higher stratum latent construct, wherein the lower stratum of internally consistent facets are used as manifest indicators of the higher stratum, or higher order, construct. Advantages of this approach include, but are not limited to, keeping the multidimensional nature of the construct explicit, and allowing the unique component of a facet to relate to other constructs in the model (e.g., using the "unique effect" approach described by Hoyle & Smith, 1994).

The second, domain-representative approach attempts to account for multidimensionality by creating parcels that encompass not only the common variance (as in the internal-consistency approach), but also the reliable unique facets of the multiple dimensions. With this method, parcels are created by joining items

from different facets into item sets. For example, the first parcel may consist of the sum of A_1, B_1, and C_1, the second parcel would be the sum of A_2, B_2, and C_2, and, finally, the third parcel would reflect A_3, B_3, and C_3. In this manner, each parcel reflects all of the facets (or dimensions) present within the set of indicators.

Consider once again the extroversion construct, as assessed using the NEO–PI (Costa & McCrae, 1992). If one were to construct internally consistent parcels for extroversion, one would create six parcels, one each for gregariousness, assertiveness, activity, excitement seeking, positive emotions, and warmth; each parcel would consist of the sum of all items for a given facet (e.g., all items from the gregarious facet). However, to construct domain representative parcels, any number of parcels could be constructed—three parcels, four parcels, or more. Under this approach, one should ensure that each parcel contains item content from each of the six facets. Each parcel should have at least one item from each of the six facets of extroversion, and each facet would preferably be present in each parcel to the same extent.

In comparing the merits of internally consistent and domain representative parcels, Kishton and Widaman (1994) reported analyses of data on three scales, one of which was a measure of internal–external locus of control. This locus of control scale had three identifiable dimensions, consistent with over 20 years of factor analytic research on locus of control. When internally consistent indicators of locus of control were included in a model, the resulting model had several problems, including both highly unstable and unacceptable parameter estimates. In contrast, use of domain representative parcels resulted in stable and acceptable estimates of all parameters. Although some may consider the domain representative parcels as representing confounded indicators, the better stability and fit of the model employing the domain representative parcels offers compelling evidence of their utility in certain situations.

WEIGHING THE MERITS

Throughout our discussion of the pros and cons of parceling, we have mentioned both merits and cautions along the way. In this section, we assemble the common general themes for a more direct contrast. We turn first to the question of dimensionality.

Nearly all of the research and literature related to parceling supports the position that the dimensional nature of a measured construct can have a serious impact on the accuracy and validity of various parceling techniques. Moreover, numerous writers have suggested that only under conditions of unidimensionality should parceling be considered (Bandalos & Finney, 2001). We concur that parceling can be particularly effective when items from a unidimensional scale are parceled. For example, some reaction time studies rely on large numbers of trials that theoretically

tap into the same underlying process or phenomenon. To estimate a latent construct with each individual trial as an indicator would clearly pose numerous difficulties. In such situations and if the dimensional structure is clear, parceling would be warranted, if not essential (see, e.g., Cunningham, Preacher, & Banaji, 2001).

The largest threats to the validity of parceling are model misspecification in general and a specific form of misspecification, multidimensionality. If parcels can obscure invalid assumptions in a model, then a researcher can never be assured that a proposed model is legitimate or correctly specified. For example, an item may load on both a depression factor and an anxiety factor, whereas a parcel can be created using the item such that the parcel loads solely on the depression construct. In such a model, some may argue that the item information is misspecified—variance associated with anxiety in the item is not captured in the model. Although such arguments appear to undermine the usefulness of parceling, we suggest that these arguments may be valid only when one is interested in the items themselves.

One can take one of two approaches to modeling latent variables. The first approach attempts to understand fully the relations among items. If this level of analysis were the primary goal, then missing a double loading or correlated residual at the item level would reflect a failure to understand fully the pattern of observed data. A second approach focuses principally on the relations among latent variables. From this perspective, item indicators are merely tools that allow one to build a measurement model for a desired latent construct. Once built, the item indicators become less consequential. With such an approach, if a dual loading were eliminated through aggregating items in order to specify a clean latent construct, then the goals of the researcher are realized through parceling, not hindered by it. The dual loading is unimportant and can be effectively minimized during an initial construct-building phase. The same logic applies to correlated residuals. If items are only building blocks, estimating the additional shared relationships is unimportant to the theory building in latent space. Eliminating the residual through parceling is as effective as explicitly correlating the residual.

The pitfalls of parceling are most evident when one seeks to understand the exact relations among the individual items comprising the measured variables. If the exact relations among items are the focus of the modeling, one should not parcel. On the other hand, if the relations among constructs are of focal interest, parceling is more strongly warranted. However, before one chooses to employ a particular technique, the data must be commensurate with the applied technique. For example, the problems associated with a misspecified model can be circumvented if prior analyses are performed to establish the factorial structure of the items that are to be modeled.

A further consideration relevant to the choice to parcel or not to parcel involves the goals of a study and the resulting goals of the measurement process. As Campbell and Fiske (1959) observed over 40 years ago, one researcher's trait construct may represent another researcher's method confound. To make headway against

this conundrum, the target constructs of a research study must be defined clearly by the investigator, and measurement operations should follow directly and unambiguously from these construct definitions. In developing measurement operations, the researcher should realize that item homogeneity is, to some extent, arbitrary. The goal should be to measure the construct of interest in as adequate a way as possible.

The arbitrariness of homogeneity can be illustrated well using the research of Spearman (1927), the first and strongest proponent of general intelligence, or *g*. In his empirical studies, Spearman typically chose tests such that the battery of measures contained at most one test of verbal comprehension, one of spatial skills, one of reasoning, and so on. Given the purposeful selection of measures, which ensured only modest overlap in the skills tapped, finding that a single factor is sufficient to explain the covariations among the measures is not surprising. Spearman labeled his single factor "g" because a single dimension could easily account for the relations among the tests.[2]

In contrast to Spearman's (1927) approach to indicator selection, Thurstone (1938; Thurstone & Thurstone, 1941) typically sought to assemble a large battery that included several tests of each given type such as several of verbal skills, several of spatial skills, and so on. Given this selection of indicators, finding that several factors (e.g., verbal, spatial, numerical, etc.) were required to model the relations among the tests was, once again, unsurprising. By constructing parcels, however, one could have restored the unidimensionality of measures in the Thurstone battery to reveal only the g component of the ability battery. For example, one could sum all verbal skill tests into a single verbal indicator, all spatial tests into a single spatial indicator, and so on. The upshot of this approach would be the restoration of a single, unidimensional latent variable model.

The measurement approaches of Spearman and Thurstone exemplifies the interplay between a researcher's goals for a study and the coordinated measurement operations that subserve that goal. If one's goal were to study g then Spearman's approach is quite reasonable and few would argue that the structure was arbitrarily manufactured. On the other hand, with the same goal in mind, one could follow Thurstone's measurement strategy to assemble a comprehensive battery of tests with several tests from each facet, but then use parceling to restore the desired unidimensional structure. Either approach would lead to a set of measured variables that would fit a unidimensional factor structure; that is, both would lead to the good fit of a measurement model with a single latent variable—provided the

---

[2]Occasionally, Spearman (1927) included several tests of a common type, such as several tests of verbal skills, but these isolated studies included only tests of the given type (e.g., all were tests of verbal skills). Once again, only a single factor was required to represent the relations among tests in such a battery, which he persisted in labeling *g*. Arguably, a more accurate label for this factor would be more specific, such as verbal comprehension. Unfortunately, such an assertion cannot be verified given that only one dimension was selected for in his analysis.

tests selected under the Thurstone method were parceled appropriately. If one's goal were to represent the structure among all single indicators, then the Spearman and Thurstone approaches to battery construction would lead to very different outcomes. Due to the fact that Spearman selected only a single test from each facet of the domain, he virtually ensured the appearance of only a single factor. Although a Thurstonian approach, in the absence of parceling, would require multiple factors, a researcher would have greater flexibility for modeling the relations among the measures by careful and informed use of parceling.

This discussion of parceling in the ability domain has clear implications for the parceling controversy that has arisen in personality realm. The primary implication is that both proponents and opponents of parceling are correct some of the time, and neither is correct all of the time. If the goal of an investigator is to model effects of a latent variable at a given level of generality (analogous to general intelligence), then appropriate selection of scales or parceling of items can minimize or cancel out the effects of nuisance factors at a lower level of generality (analogous to verbal comprehension, spatial ability, etc.). In such situations, parceling is warranted. If the investigator's goal is to represent the dimensionality of the measurement space at the level of the individual tests or items, then the minimizing of lower-level effects would tend to obscure precisely the effects that the investigator intends to study. In situations of this type, parceling is contraindicated. Careful delineation of the goals of a study is clearly the paramount issue. Once the goals are carefully laid out, the decision to parcel or not to parcel would be dictated by the goals and the nature of the measurements obtained. A researcher's goals and the measurement operations needed to attain these goals supercede either doctrinaire stance regarding the decision to parcel or not to parcel. In the end, two clear conclusions can be drawn from our review of the issues. On the one hand, the use of parceling techniques cannot be dismissed out of hand. On the other, the unconsidered use of parceling techniques is never warranted.

ACKNOWLEDGMENTS

REFERENCES

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

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1,* 35–67.

Bandalos, D. L. (1997). Assessing sources of error in structural equation models: The effects of sample size, reliability, and model misspecifications. *Structural Equation Modeling, 4,* 177–192.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cattell, R. B. (1956). Validation and intensification of the sixteen personality factor questionnaire. *Journal of Clinical Psychology, 12,* 205–214.

Cattell, R. B. (1961). Theory of situational, instrument, second order, and refraction factors in personality structure research. *Psychological Bulletin, 58,* 160–174.

Cattell, R. B., & Burdsal, C. A., Jr. (1975). The radial parceling double factoring design: A solution to the item-vs.-parcel controversy. *Multivariate Behavioral Research, 10,* 165–179.

Costa, P., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitudes measures: Consistency, stability, and convergent validity. *Psychological Science, 12,* 163–170.

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

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.

Hau, K-T., & Marsh, H. W. (2001). *The use of item parcels in structural equation modeling: Nonnomral data and small sample sizes.* Manuscript submitted for publication.

Holahan, C. J., & Moos, R. H. (1994). Life stressors, person and social resources, and depression: A 4-year structural model. *Journal of Abnormal Psychology, 100,* 31–38.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research questions as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology, 62,* 158–176.

Karasawa, M., Little, T. D., Miyashita, T., Mashima, M., & Azuma, H. (1997). Japanese children's action-control beliefs about school performance. *International Journal of Behavioral Development, 20,* 405–423.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757–765.

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

Little, T. D., 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.

Little, T. D., Oettingen, G., & Baltes, P. B. (1995). *The revised control, agency, and means-ends interview (CAMI): A multi-cultural validity assessment using mean and covariance (MACS) analyses (Materialen aus der Bildungsforschung, No. 49).* Berlin: Max Planck Institute.

Little, T. D., & Wanner, B. (1997). *The Multi-CAM: A multidimensional instrument to assess children's action-control motives, beliefs, and behaviors (Materialen aus der Bildungsforschung, No. 59).* Berlin: Max Planck Institute.

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

Marsh, H. W., Hau, K. T., 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.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107–117.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38,* 30–38.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

SEMNET. (2001). Structural equation modeling discussion network at Listserv@BAMA.VA.edu

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology, 71,* 549–570.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Tanaka, J. S., & Huba, G. J. (1987). Assessing the stability of depression in college students. *Multivariate Behavioral Research, 22,* 5–19.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence* (Psychometric Monographs, No. 2). Chicago: University of Chicago Press.