Psychological Methods
1999, Vol. 4, No. 2, 192–211

1082-989X/99/$3.00

On Selecting Indicators for Multivariate Measurement and Modeling With Latent Variables: When "Good" Indicators Are Bad and "Bad" Indicators Are Good

Todd D. Little
Yale University

Ulman Lindenberger
Max Planck Institute for Human Development

John R. Nesselroade
University of Virginia

Selecting indicators is as important for the generalizability of research designs as selecting persons or occasions of measurement. Elaborating on the extant knowledge base regarding indicator selection, the authors examine selection influences on the validity and reliability of multivariate representations. A simulation that systematically varied 4 key dimensions of indicator selection was used to investigate their effects on the fidelity of construct representations and the relative ability of exploratory and confirmatory analyses to recover within- and between-construct information. Confirmatory analyses, for example, yielded valid and unbiased estimates of the relations between constructs, even under conditions of very low internal consistency. Design, procedural, and analysis recommendations based on an expanded taxonomy of indicator selection and the simulation results are provided.

Designing empirical research should oblige investigators to attend explicitly to the many measurement attributes of their expected data. Cattell (1952, 1996a), for example, explicated this point by identifying up to 10 dimensions that characterize the "data box"—his heuristic for distinguishing among possible data configurations in covariation designs. Of these dimensions, the most common and familiar attributes are persons, variables, and occasions of measurement (i.e., a datum is the score of a specific person on a given variable at a particular occasion). Given that designing an empirical study requires an investigator to select by one means or another how many and which persons, variables, and occasions will define the projected data set, these selection decisions impinge directly on the quality of research designs and the value of the results. Cronbach, Gleser, Nanda, and Rajaratnam (1972), for example, formalized many implications of these issues within the general rubric of measurement and generalizability, as did Campbell and Stanley (1963) in their discussion of experimental and quasi-experimental designs.

Given that measurement is, to a considerable degree, a selection process, we find it surprising that work on selection has focused more on persons and occasions than on the selection of indicators. For example, models of selection and selection effects (Aitken, 1934; Berk, 1983; Lawley, 1943; Lawley & Maxwell, 1971; Linn, 1973, 1983; Meredith, 1964,

The contributions of the authors to this article were equivalent.

We are very grateful to Paul Baltes, the other members of the Max Planck Institute for Human Development's Center for Lifespan Psychology, and to several colleagues, including Richard Ryan, Michael Shanahan, and Rolf Steyer, for their many helpful discussions and comments. We also thank Werner Scholtysik and Wolfgang Assmann for their computer resource management services.

Correspondence concerning this article should be addressed either to Todd D. Little, Department of Psychology, Yale University, P.O. Box 208205, New Haven, Connecticut 06520-8205 (electronic mail may be sent to todd.little@yale.edu); to Ulman Lindenberger, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany (electronic mail may be sent to lindenberger@mpib-berlin.mpg.de); or to John R. Nesselroade, Department of Psychology, University of Virginia, Charlottesville, Virginia 22903-2477 (electronic mail may be sent to JRN8Z@virginia.edu).

1993; Pearson, 1903; Poon, Krauss, & Bowles, 1984) as well as conceptions of generalizability (Cronbach et al., 1972) have alerted us to the treachery of nonrepresentative person sampling. Similarly, given certain information and assumptions, the concept of statistical power informs the choice of sample size (J. Cohen, 1992; MacCallum, Browne, & Sugawara, 1996). Developmental researchers, too, have articulated many issues and potential pitfalls in selecting occasions of measurement (Baltes, Reese, & Nesselroade, 1977; McArdle & Woodcock, 1996). On the other hand, researchers know less about how to select particular indicators (manifest variables) to represent a given construct (latent variable). Although germane notions of validity obviously apply (Campbell & Fiske, 1959; Gulliksen, 1950; Messick, 1982; Ozer, 1986), few formal guidelines have been offered that can inform one's choices of how many or which indicators to use. Instead, indicator selection has typically relied on informal or intuitive reasoning and historical precedent.

In our view, the relative lack of a formal framework for guiding such decisions introduces an unsatisfying ambiguity into research designs involving latent variables. In an effort to clarify some of the issues and to shed further light on the directions from which answers might eventually come, we operationalize four key aspects of selecting indicators to represent psychological constructs and examine their impact on research outcomes. In particular, we focus on three primary goals in this article. First, from both a generalized selection orientation (Nesselroade & Jones, 1991) and a domain-sampling perspective (Nunnally, 1978; Nunnally & Bernstein, 1994), we identify four specific dimensions of indicator selection that can affect the fidelity of construct representations in multivariate research. Second, by means of a Monte Carlo simulation, we systematically investigate their influences on the relative efficacy of various analytic techniques (exploratory, confirmatory, raw data) to recover information about constructs that is inherent in the relations among indicators. Third, on the basis of the simulation results, we provide some design recommendations and cautions. Given the clear advantages and current popularity of latent-variable approaches (Bentler, 1980; Little, 1997; McArdle, 1996), we believe that (re)directing attention to such basic measurement concerns will strengthen behavioral and social science research involving latent variables.

Before beginning our discussion of indicator selection, however, we must emphasize that at least two types of indicators and constructs can be distinguished—effect indicators of common constructs versus cause indicators of emergent constructs (see, e.g., Bollen, 1989; Bollen & Lennox, 1991; P. Cohen, Cohen, Teresi, Marchi, & Velez, 1990; MacCallum & Browne, 1993). With effect indicators, values on manifest variables are presumed to be caused by one's standing on one or more latent variables (i.e., the causal direction is from construct to indicator). With cause indicators, changes on one or more manifest variables are presumed to lead to changes in one's standing on the emergent construct (i.e., the causal direction is from indicator to construct). With this distinction in mind, the focus of our study is squarely on effect indicators of common constructs. Throughout our discussions, therefore, we use the terms indicator and manifest variable to refer broadly to any observable measurement attribute (e.g., items, behaviors, responses) of the effect-indicator type, and we use the terms construct and latent variable to refer broadly to any underlying hypothetical abstractions of the common type.

In a similar vein, we must also emphasize that the domain-sampling metaphor, which we rely on below to illustrate the various aspects of indicator selection, is based on an effect-indicator model of measurement. Although the domain-sampling metaphor is but one of a number of metaphors that can be used to think about measurement issues (Nunnally, 1978; Nunnally & Bernstein, 1994), it and various other metaphors are easily used to derive the central formulae of classical test theory. In other words, these metaphors and the fundamental logic of classical test theory are primarily concerned with effect indicators of constructs. Moreover, the bulk of the analysis machinery that is available to researchers in the social and behavioral sciences is, broadly speaking, geared toward explicating relationships in data that derive from such a model. Given the substantial prevalence of effect indicators of constructs in most areas of the behavioral sciences, this restriction would seem reasonable and leaves room for future discussions of cause indicators.

Selecting Indicators in Psychological Research

Brief Historical Background

Selecting variables in psychological research has been a long-standing concern, even though the volume of attention has been relatively low. For example, the importance of variable selection can be seen in

Thurstone's (1938; Thurstone & Thurstone, 1941) writings on primary mental abilities. He described spending as much time on developing quality measures as on the factor analyses that were conducted to evaluate them (at a time when only laborious hand calculations were possible). Within the context of research on the personality sphere, Cattell (1952) examined the matter of selecting variables at great length. His proposals for defining a *content domain* and sampling from it remain influential (Goldberg, 1993; Nunnally, 1978). Humphreys (1962), too, discussed many key issues of selecting indicators and proposed a rationale for constructing measures that features the concept of *controlled heterogeneity*—an optimal (but not necessarily maximal) level of the characteristic reflected in coefficient alpha (Cronbach, 1951). At about the same time, Campbell and Fiske (1959) addressed questions about the relations among manifest and latent variables and initiated a line of inquiry concerning *convergent* and *discriminant* validity that continues to the present (Widaman, 1992). For instance, Campbell and Fiske's notion of *method variance,* along with Cattell's (1961) *instrument-factor* conception, has been brought directly into contemporary work on trait–state distinctions (Dumenci & Windle, 1996; Steyer, Ferring, & Schmitt, 1992). In addition, Kaiser and his colleagues (Cerny & Kaiser, 1977; Kaiser, 1970; Kaiser & Rice, 1974) introduced the idea of the *sampling adequacy* of indicators—the empirical index of which is still integrated into various statistical approaches (e.g., SAS Institute, 1990).

These concepts share the concerns that the precise location of indicators and constructs in multivariate space is unknown and that their interrelations must be inferred. In addition, these ideas highlight the fact that the negative consequences of this uncertainty are much alleviated by measuring more than one or two variables to represent a given construct—the multivariate approach (Baltes & Nesselroade, 1973; Cattell, 1966b). From this viewpoint, various relations among sets of manifest variables are used to make inferences about the relative locations of the respective constructs in multivariate space. In other words, estimates, which are empirically derived from the selected data and potentially fallible, are used to intimate the true multivariate relations.[1] To some degree, information such as internal-consistency measures and factor-loading patterns can be used to judge the adequacy with which sets of indicators (i.e., items, behaviors, responses) represent hypothetical abstractions such as aggression, intelligence, self-efficacy, and the like. However, as we describe in more detail below, overreliance on such information (e.g., assuming that more highly intercorrelated variables lead to better construct representations than do less intercorrelated variables) can be misleading under specifiable selection conditions, especially when one has relatively poor knowledge about the precise location of a construct's centroid in multivariate space.

A Vector-Space Representation of Multivariate Space

In addition to a generalized selection orientation, we adopt a domain-sampling model as our substantive heuristic. From a domain-sampling viewpoint, a given construct has a broad universe of possible items, behaviors, and responses that can serve as its observable markers or indicators (Nunnally, 1978). Moreover, the region in multivariate space inhabited by indicators of a given construct can be quite small (e.g., numerical facility) or relatively large (e.g., intelligence). Even though we rely on the domain-sampling model to illustrate our data generation schema (see below), the implications for measurement are not limited to constructs for which the universe of possible items is "infinite." As long as some number of indicators can be selected to represent a construct (even if that number is rather small), the measurement implications derived from our taxonomy and simulation hold for constructs represented by effect indicators. To illustrate these ideas and to introduce concepts that are central to our selection-based simulation, we use a vector-space, or geometric, representation of multivariate space (see Gorsuch, 1988; Harman, 1967).

To avoid unnecessary complexity, we restrict our example to three geometric dimensions and start with the notion of the unit sphere—a ball-like space wherein the relations among all points within the sphere can be measured in a standardized metric. Within this three-dimensional sphere, which is presented schematically in Figure 1A, any observable or manifest variable (potential indicator) is represented as a directed line segment, or vector. All such indicator vectors emanate from the center of the sphere (the ori-

[1] Throughout our discussion of indicator selection, we use the term "true" to distinguish between real relations among the various multivariate features (i.e., in the population sense) and the estimated, or inferred, relations (i.e., in the sampled sense).

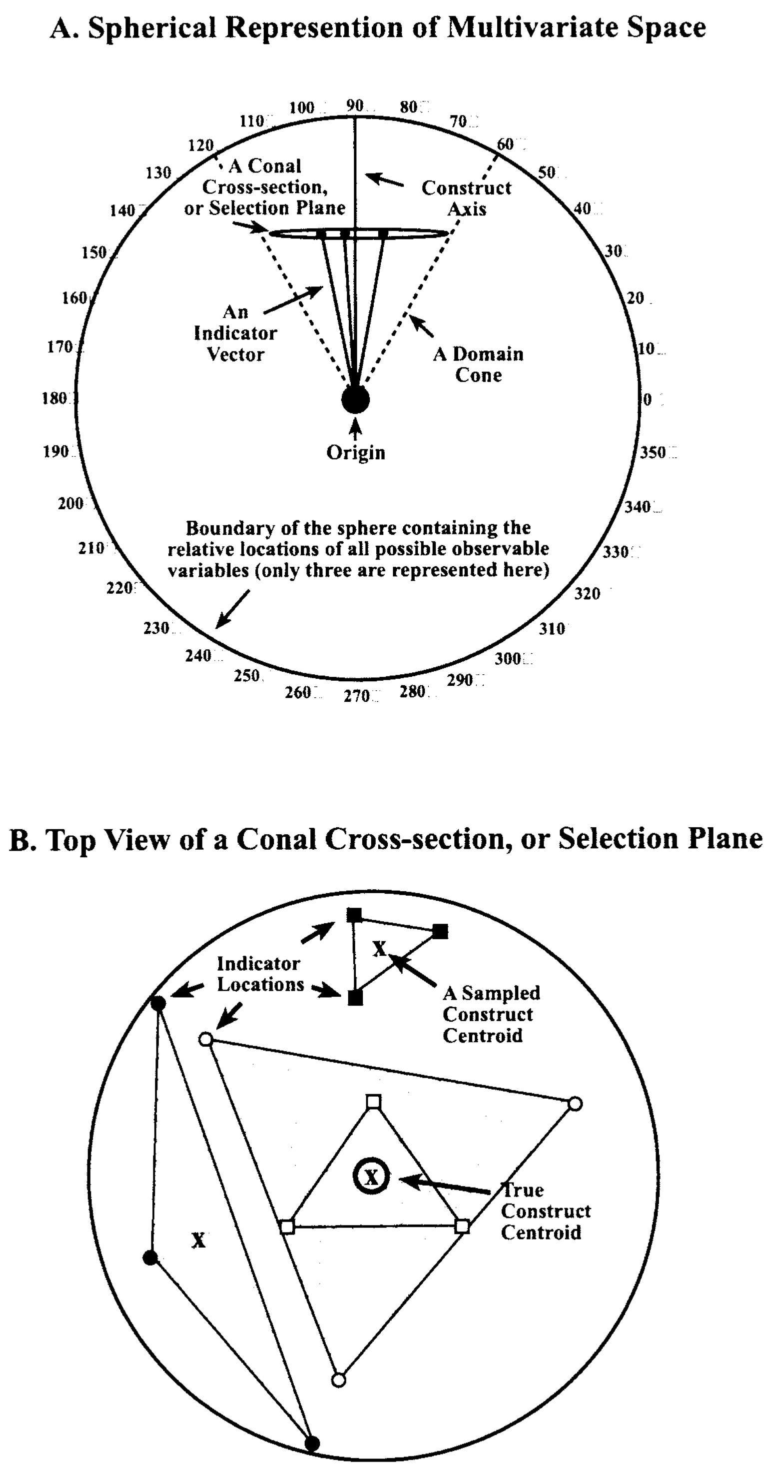

Figure 1. The unit sphere and a geometric representation of four possible three-variable selections of indicators (manifest variables) to mark a construct's true domain centroid.

gin), extend toward the surface, and terminate at a specific point in multivariate space. A construct (latent variable, factor) is also represented by a vector, but, because this vector emanates from the center of the sphere and, by definition, extends to its surface, we refer to it as a construct axis, or centroid.

Within this representation, a construct domain can be represented as a cone of unspecified volume surrounding a given construct's axis with the cone's apex falling at the origin and its base intersecting the surface of the sphere in a circle.[2] One such possible domain cone is represented in Figure 1A. In geometric terms, the length of an indicator vector is the square root of its reliability (in the sense of classical test theory). The correlation between any two indicators is a function of their angular distance, weighted by their respective degrees of reliability (the cosine of the angle between the two indicators multiplied by the respective lengths of their vectors).

Metaphorically speaking, the distribution of indicators throughout the volume of a cone symbolizes the array of possible indicators that one may select in designing an empirical study. Other things being equal, the closer a selected indicator falls to the construct axis in the center of a cone (i.e., the construct's domain centroid), the more veridical is its construct representation. As presented in Figure 1B, the four different three-variable constellations vary in their closeness to a construct's true location in multivariate space (i.e., a construct's centroid), as well as in their degree of intercorrelation, depending on the relative locations of the indicators. On the one hand, the sampled centroid of three variables with relatively low intercorrelations (the widely spread unfilled circles in Figure 1B) can yield a very good approximation of the centroid to the degree that the three indicators diverge from the construct axis in a balanced manner such as in equidistant and equidivergent directions (cf. the filled circles in Figure 1B). On the other hand, the sampled centroid of three highly correlated variables (closed grouped) may yield a biased, or off-center, representation of the construct to the degree that their sampled centroid diverges from the true construct centroid (cf. the filled squares, reflecting a biased representation, with the unfilled squares as shown in Figure 1B).

Figure 1B illustrates that knowledge of a construct's location in multivariate space is generally imperfect. Moreover, this knowledge can vary considerably across different research domains, depending on the quality of prior theorizing and empirical work. For instance, the location of the centroid of intellectual functioning in multivariate space is, by some accounts, relatively well defined (Carroll, 1993; Marshalek, Lohman, & Snow, 1983). As we emphasize in more detail below, the amount of knowledge about construct locations has important implications for optimal construct representations because some of the relevant dimensions of indicator selection interact with the degree of certainty with which researchers can locate their constructs in multivariate space.

Four Dimensions of Indicator Selection

To provide an operational framework that builds on and integrates many of these issues, we define and discuss four dimensions of indicator selection that can affect the fidelity, or quality, of construct representations: (a) centroid distance, (b) number of indicators, (c) selection communality, and (d) selection diversity. In our view, these four dimensions provide a useful way to conceptualize and understand how indicator selection effects can arise.

Centroid Distance

Centroid distance is defined as the true distance, or correlation, between any two constructs, which, in our vector-space representation, corresponds to the angular distance between any two construct centroids (i.e., the cosine of the angle between the axes representing the construct centroids). When the distance between two centroids is high, their correlation is low, and indicators of the different constructs are likely to populate clearly separable regions of multivariate space. When the distance between two centroids is low, their correlation is high, and indicators of different constructs are likely to populate adjacent, or even overlapping, regions of multivariate space. This latter situation, depending on the nature of the indicator selections associated with each construct, may make it difficult to properly recover the constructs' locations in multivariate space and, therefore, their true correlation.

---

[2] Although the hypothetical universe of potential indicators reflects an unspecified volume within a domain, both theory and operationalization processes, by focusing on the centroid of constructs, suggest boundary conditions that give shape to a domain cone.

Number of Indicators

The second selection dimension is the *number of indicators*. Classical measurement theory indicates that, all other things being equal, more items lead to better construct representations. Stated simply, as more and more indicators are selected, the centroid of the sampled indicators will lie closer to the true centroid of a domain. Under unbiased (e.g., random) selection conditions, if enough variables are selected, the expected value of a sampled centroid will approximate the construct's true centroid because it is unlikely that all indicators are off-center in the same direction. For numerous practical reasons, however (e.g., the practical cost–benefit trade-offs among expedience, breadth of coverage, and parsimony vs. maximal accuracy and consistency), finding an optimal rather than a maximal number of indicators is a desired feature in research design. As a consequence of this desire for a sufficient yet small number of indicators of a given construct, the importance and relative influence of the next two selection dimensions become even more relevant.

Selection Communality and Selection Diversity

The next two dimensions of indicator selection are the degree of *selection communality* and the degree of *selection diversity*. These concepts and the distinction between them are illustrated in Figure 2. We define these two dimensions in relation to the concept of a selection plane. As previously defined (see Figure 1A), a selection plane is the circular cross section of a possible domain cone centered on the centroid (construct axis) at a given altitude or height. In our framework, a selection plane is a flat circular area at a specific location in multivariate space from which indicators can be selected (see Figures 1 and 2).

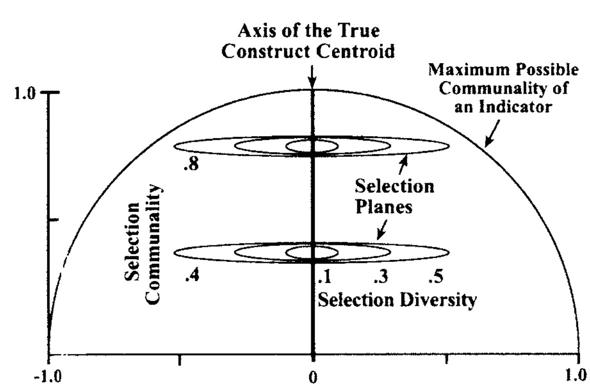

Figure 2. Geometric representation of selection planes defined by two of the dimensions in selecting indicators (manifest variables) of a given construct (latent variable): selection communality and selection diversity.

The location of a selection plane is determined by its values on two dimensions. The first dimension is the height or altitude of the selection plane on the construct axis and reflects its degree of selection communality. Here, selection planes at higher altitudes represent indicators that have high construct-related variance (e.g., scales based on many items or aggregated responses), whereas selection planes at lower altitudes represent indicators that have low construct-related variance (e.g., a single item or response). The second defining dimension of a selection plane is its width or circumference and reflects its degree of selection diversity (i.e., the diameter of a selection plane centered on the construct axis as shown in Figure 2; see also Cattell & Tsujioka, 1964). Here, selection planes of narrower diversity (i.e., of smaller circumference) represent indicators with relatively small amounts of specific (but reliable) variance, whereas selection planes of broader diversity (i.e., of larger circumference) represent indicators with potentially greater amounts of specific variance (i.e., complex, heterogeneous indicators).

Given that the reliability of any indicator cannot exceed 1.0, these defining dimensions of a given selection plane (i.e., its communality and diversity) are not fully independent—the higher a selection plane's communality (altitude), the narrower its maximal diversity (circumference). This necessary, albeit peripheral, constraint on these two defining dimensions of a given selection plane is represented by the outer semicircular arc in Figure 2 and reflects the maximum possible width of a selection plane at a given level of selection communality. In particular, the atypical extremes of selection are constrained such that when the communality of a selection plane is 1.0, the degree of diversity can only be 0 (completely isomorphic indicators). The maximum diversity of a selection plane increases rapidly as selection communality decreases such that, again in the extreme case, when selection diversity is 1.0, selection communality must be 0 (candidate indicators are perpendicular to the construct axis).

Taken to the logical extreme, all manifest variables are candidate indicants of any given construct, allowing for reflection, or reverse coding, of negatively related indicators. However, as mentioned, both practical and theoretical considerations such as reasonable face validity, adequate construct variability, discrimination among constructs, and specific operationalizations limit typical measurement designs to only a subset of possible indicators. For example, although large levels of selection diversity are possible and may actually occur in practice, very wide diversity becomes difficult to interpret. In extreme cases, the amount of common-construct variance would likely be very small and perhaps nonsignificant, or the indicators would be too highly related with other constructs in multivariate space to disentangle the construct relations that are being pursued.

Some illustrative values of selection communality and selection diversity are depicted in Figure 2. For each level of selection communality (.4 and .8), three levels of selection diversity, ranging from quite limited (.1 diversity) to quite sizable (.5 diversity), are represented. Note that indicators whose termini populate a given selection plane differ in terms of their closeness to each other and to a domain's true centroid (see Figure 1B). At a given level of selection communality, indicators from tighter planes (less diverse and therefore reflecting less specificity; i.e., reliable indicator-specific variance) have less variability in their relations with one another and tend to be closer to the construct centroid than do indicators from wider spheres (greater diversity). Given that wider diversity reflects greater complexity in the variance composition of an indicator, indicators from diverse-selection spheres would have a greater potential to share variance with other indicators of other constructs.

Together, selection communality and selection diversity determine the nature of the construct properties of a selected indicator (i.e., indicator communality, specificity, and reliability). *Indicator communality* refers to the degree of reliable construct variance, or communality, shared between an indicator and the centroid of a construct (i.e., true construct-related variance). In other words, the selection communality of the selection plane determines the magnitude of a given indicator's loading under unbiased selection. *Indicator specificity* refers to the degree of homo- or heterogeneity of a given indicator (i.e., its distance from a construct centroid). Here, under unbiased selection, selection diversity determines the reliable component of a given indicator's unique or residual variance. Finally, these two properties, together, comprise the total reliability of an indicator. That is, *indicator reliability* is determined by an indicator's degree of communality plus its degree of specificity.[3]

In practice, the reliability of an indicator is inferred from such indices as item–scale correlations and factor-space communality estimates. A problem with interpreting such indices, however, is that we do not know the location of the true centroid, and therefore we cannot be sure of the true composition of an indicators' reliability. For instance, we do not know whether an indicator's reliability results from a combination of either a high level of communality coupled with low specificity or of a low level of communality coupled with high specificity. Stated another way, analysis techniques typically attempt to decompose an indicator's variance into a construct-common component and a unique component. The unique component is assumed to contain both indicator-specific variance and unreliable variance (both of which are assumed to be uncorrelated with other variance components among the multivariate indicators of a given analysis). However, such decompositions and inferences yield inaccurate information about the nature of a construct if the selection of indicators is biased (see Figure 1B).

An Illustrative Example

Figure 3 further illustrates the four dimensions of indicator selection and the operational nature of our Monte Carlo simulation. In the top panel of Figure 3, two constructs (A and B) are shown. To simplify this example, the two constructs are defined to be orthogonal to each other (uncorrelated), and they are standardized to reflect relations that are of unit length (i.e., correlational relations are depicted between the constructs and their possible indicators).[4]

In Figure 3, we depict one selection plane for each construct at a high level of selection communality and a moderate level of selection diversity. The circular representation of a possible selection plane symbol-

[3] Referring back to Figure 1A, consider any indicator vector in the domain of the construct (and thus falling somewhere inside the cone). That vector results from two component pieces: (a) the length of its projection on the domain centroid and (b) the distance of its endpoint from the domain centroid. An indicator's reliable variance is partitioned between its communality with the centroid and its heterogeneity, or diversity, with respect to that centroid. This latter is analogous to specificity in common-factor model terminology. Using the Pythagorean theorem, one can say that an indicator's reliability is the square root of its squared communality plus its squared specificity.

[4] Note that all aspects of this simulation hold for covariance relations as well.

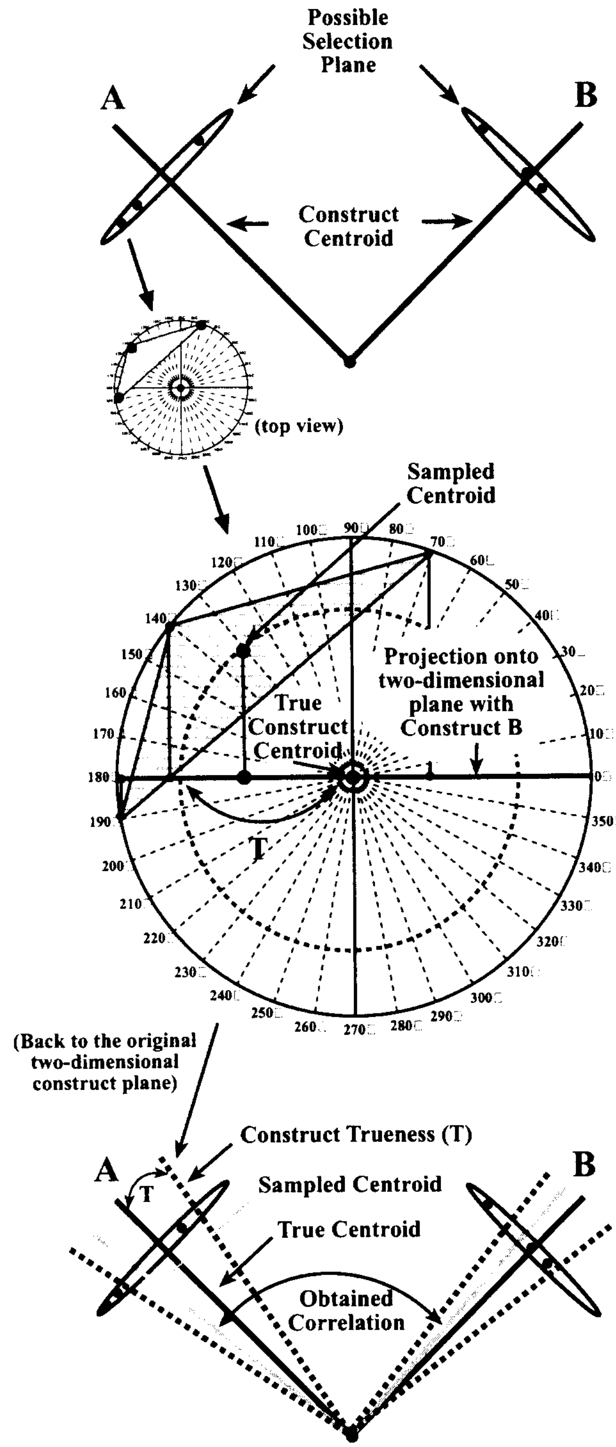

Figure 3. Geometric representation of conceptual aspects of selecting indicators of constructs. (Note. *T* represents the construct trueness of the selections and is defined as the largest possible angular distance between the centroid of the sampled indicators and the true centroid of the desired construct that a sampled centroid may have after selection.)

izes the radial array of potential indicators that define a construct at a given level of selection communality and selection diversity. In the middle panel of Figure 3, one possible selection of three indicators for Construct A is shown. In this example, the selected indicators define a sampled centroid that departs from the true construct centroid by some amount due to the inherent variability associated with any one selection of indicators. Finally, the selected indicators and the sampled centroid are projected back onto the two-dimensional construct plane represented in the bottom panel of Figure 3.

Consistent with the propositions of classical test theory, the four selection dimensions (centroid distance, number of indicators, selection communality, and selection diversity) and their influence on the true construct composition of indicators (true indicator communality, specificity, and reliability) provide an exhaustive operational categorization of a given indicator's relations in multivariate construct space. In our view, attending to these four dimensions as explicitly as possible when choosing and operationalizing the possible indicators of the constructs in a given study should yield a broader appreciation for the potential limits to the accuracy and generalizability of a study's results and thereby enhance the overall quality of research.

Modeling the Relations Among Selected Indicators

In addition to the general measurement concerns detailed above, the ambiguities associated with these selection dimensions can also affect the efficacy of various analytic techniques in recovering the inherent information about the underlying multivariate relationships (i.e., indicator-to-indicator, indicator-to-construct, and construct-to-construct). With the amazing growth of structural modeling applications in the past few decades, the questions and issues of indicator selection are particularly dramatized because of their strict dependence on multiple indicators to define a given construct (e.g., Bentler, 1980; Bollen, 1989; Jöreskog & Sörbom, 1989; McArdle, 1996; McArdle & McDonald, 1984; Mulaik, 1988). Therefore, to shed some additional light on the relative virtues of exploratory and confirmatory approaches, we also address the question: Do exploratory or confirmatory analytic techniques more accurately recover information concerning the true relationships among latent variables? More precisely, we examined whether exploratory and confirmatory analytic techniques systematically differ in their ability to yield accurate (unbiased) and efficient estimates of the true association between constructs.

Method

As mentioned, we focus our simulation on the common-factor model, with important roots embedded in classical test theory, for representing the relations among indicators and constructs. To simulate a broad array of research-design situations, we generated 60,000 sample correlation matrices reflecting systematic variations of (a) centroid distance, (b) number of indicators, (c) selection communality, and (d) selection diversity. For ease of communication, we included only two constructs, which varied in their centroid distance, or true correlation, over four levels (0, .3, .5, and .7). The number of indicators for the constructs varied over five levels (2, 3, 4, 5, and 6). Although using only two indicators per construct is problematic, such a situation is often found in empirical research; therefore, we included it to examine its behavior in the broader context of our simulation.

In addition to these two more transparent features of our simulation, we varied the magnitude of each selection plane's communality (i.e., the altitude, or location of the selection plane) over three levels (.4, .6, and .8). To vary the degree of selection communality systematically, we selected the multiple indicators at the same selection communality. That is, we constrained the degree of indicator-to-construct communality to be a constant across all indicators whose termini populate a given selection plane (i.e., the altitude, or location of the plane, from which an indicator is selected is the same for all indicators, regardless of the degree of selection diversity). However, the results and implications of our simulation do not depend on such constraints (i.e., they generalize across mixed communality conditions). Finally, at each level of selection communality, we also varied the maximum degree of selection diversity over five levels (.1, .2, .3, .4, and .5). Specifically, we varied the degree of indicator specificity uniformly from zero to the maximum defined by the circumference of the selection plane. Thus, the several indicator vectors of a selection plane share the same selection communality but vary in their diversity. As a result, the indicators also vary in their reliability because an indicator's reliability is determined by its degree of communality plus its degree of diversity (see Footnote 3).

Combining the four varied features for the data generation phase of the simulation results in a design matrix of 300 cells. Within each cell, we generated 200 sample correlation matrices by drawing the designated number of manifest variables at random from the possible indicators of a construct's sample space.[5] At each level of selection communality, we defined the indicators' sampling space as a selection plane containing 360 possible sectors (one for each degree of arc in the circle defining a selection plane) with a continuous range of diversity bounded by 0 (perfectly accurate) to the maximum diversity of the selection plane (e.g., .1, .2, .3, .4, or .5). Each sampled correlation matrix was generated by randomly selecting points in this geometric space and calculating all possible vector distances using basic geometry. The steps and processes involved in the data generation are illustrated in Figure 3.

The first analytic target was to estimate the true correlations between the two constructs on the basis of the correlations within and between their respective indicators. To do this, we analyzed the data using both exploratory and confirmatory methods. Specifically, we used a standard confirmatory structural-equations solution (i.e., maximum-likelihood estimation using Proc CALIS; SAS Institute, 1990) wherein only the expected loadings were estimated simultaneously with the latent correlation between the two constructs. For the exploratory procedure, we report the results from an iterated maximum-likelihood estimation with an oblique Harris-Kaiser rotation (i.e., using Proc FACTOR with the HK rotation option and HKP = 0; see SAS Institute, 1990) because this factor model is both quite common and most similar to the confirmatory approach.[6] Finally, as a further comparison, we also estimated the constructs' correlation using a standard raw-data technique, namely, the Pearson correlation between the unit-weighted composite of the indicators (i.e., the cosine of the angle between the two axes defined by the unit-length geometric average of each construct's respective indicators).

We next compared the various estimates of the construct correlation with the actual construct correlation (i.e., the correlation specified in the simulation) to ascertain the influence of the manipulated design variables. The primary dependent variable across these analyses was the difference between the estimated correlation and the true population correlation. In more operational terms, the dependent variable was a measure of how precisely information about the constructs (e.g., their centroids and intercorrelation) could be recovered by analyzing relationships among their indicators chosen in various configurations.

To determine the influence of the manipulated design variables on the recovery of the true correlations, we conducted a 3 (method of analysis) × 5 (selection diversity) × 4 (number of indicators, with the two-indicator case removed) × 4 (centroid distance) × 3 (selection communality) analysis of variance (ANOVA), with a priori contrasts examining the relative differences between the methods of analysis. For ease of communication, this full ANOVA was followed up by a separate ANOVA for each method of estimation. For these analyses, negative values of the dependent variable indicate an underestimation of the true correlation, and positive values indicate an overestimation of the true correlation. Because of the tremendous power of the significant tests due to the very large sampling of data, we focused only on the pronounced and consistent effects (i.e., $p < .0001$ for each method of analysis). Finally, we also examined the variability of the estimates in evaluating the outcomes of the simulation. Although two quite different combinations of cells in our design can yield an average dependent variable value of zero, indicating an unbiased estimate of the true correlation, the design features may yield estimates of the true value with dif-

[5] The SAS/IML modules used in this simulation, as well as the results for each cell of the design, are available from Todd D. Little on request or can be downloaded directly from http://www.mpib-berlin.mpg.de/research_resources.html.

[6] In supplemental analyses, we also compared a least-squares estimator and alternative oblique rotation methods and found no fundamental differences from the results of the exploratory technique that we report. We also compared a principal-component solution and found that it performed much like the raw-data method owing to the fact that this factor model also does not account for measurement error (i.e., unreliability). For the exploratory procedures, we also specified the number of factors to be extracted to be two, rather than attempting a full comparison of rules for determining the number of factors to extract, because we found quite often in our preliminary analyses that, when the constructs were highly correlated and the selection diversity was moderate to large, only one factor was retained. Therefore, one should bear in mind that given the type of exploratory procedure that we estimated, it reflects a very close analogue to a confirmatory model with the exception that all possible loadings are estimated (i.e., a pseudoconfirmatory application of the exploratory algorithm).

ferent efficiency (i.e., greater or lesser variability across solutions).

Results

The Influence of the Four Dimensions of Indicator Selection

Before turning to the results of the comparisons of the analytic techniques to recover the correlations between the constructs, we first present a summary of the influences of the four dimensions of indicator selection. The random samples of indicators drawn in the simulation produced estimates that converged on the true correlation in the data, offering no evidence of systematic error in the nature of the selected data, $F(299, 59700) = 1.1$, $p = .11$, grand mean = 0.

Notably, however, the range of possible correlation that could emerge on any one sampling of indicators was often quite pronounced. To clarify this point, we introduce the concept of construct trueness (T) to represent the degree of possible bias (see Table 1). Referring back to Figure 3, the level of construct trueness, T, is defined as the largest possible angular distance between the centroid of the sampled indicators and the true centroid of the desired construct that a sampled centroid may have after selection (cf. Cattell & Tsujioka, 1964, on factor trueness). It is a type of validity index of the sampling bias for a given set of indicators from a selection plane of a given circumference. The values listed in Table 1 are, for three levels of selection diversity (.1, .3, and .5) and number of indicators (2–6), the mode, standard deviation, upper 95% confidence threshold, and maximum degrees of possible bias (represented as percentages: $[1 - T^2] \times 100$).

Two important aspects of the information in Table 1 are (a) the reduction in the variability in the degrees of bias that can emerge as more indicators are selected and (b) the increases in the degrees of bias that can emerge as selection diversity increases. As mentioned, these estimates of selection bias are independent of the amount of selection communality and the amount of domain overlap. These values are relevant because they reflect the degree of inherent variability that can emerge when trying to measure the centroid of a given construct, irrespective of a chosen analytic technique. These values can be used as a heuristic guide to determine the degree of diversity and number of indicators that a researcher can tolerate for a given research question. For example, if the available indicators of a construct are quite diverse, then a researcher might select more indicators to counter the effects of the indicators' diversity.

Differences in Bias and Efficiency Among the Analytic Techniques

The different methods of analysis showed substantial differences in the ability to recover the true correlation among the constructs, $F(2, 143280) = 20{,}978.5$, $p < .0001$. The a priori contrasts showed that each method differed from the others (all ps < .0001). The follow-up ANOVA results are presented in Table 2. A number of general features can be gleaned from the information in Table 2 regarding the differences in the various analytic techniques to recover the true correlation between the constructs.

Overall bias effects. As a whole, the four dimensions of indicator selection had the least influence on the analytic efficacy of the confirmatory structural modeling technique. Only 1.5% of the variance in the

Table 1
Bias and Efficiency at Specific Levels of Selection Diversity and Number of Indicators

	Low (.1) diversity				Medium (.3) diversity				High (.5) diversity			
	Bias	Efficiency			Bias	Efficiency			Bias	Efficiency		
Number of indicators	Md	*SD*	95%	Max	Md	*SD*	95%	Max	Md	*SD*	95%	Max
2	0	.16	.5	1.0	0	1.47	4.4	8.2	0	4.10	12.3	24.5
3	0	.11	.3	.8	0	1.00	3.0	7.2	0	2.72	8.2	23.0
4	0	.08	.2	.6	0	.73	2.2	5.5	0	2.09	6.3	16.5
5	0	.07	.2	.5	0	.61	1.8	4.5	0	1.66	5.0	13.6
6	0	.06	.2	.5	0	.49	1.5	3.9	0	1.44	4.3	14.0

Note. Md = mode of the trueness estimates expressed as a percentage; 95% = 95th percentile of the trueness values; Max = the maximum bias that occurred across the replications. These values are symmetric, thus they reflect the degree of possible bias, either too much or too little for a given construct. The degree of possible bias between any two constructs with the same selection conditions would be two times the tabled values. These values are independent of selection communality.

Table 2
Summary Table of F Ratios From the Analyses of Variance

Source	*df*	CFA	EFA	Raw data
A. Selection communality	2	0.9	0.3	18,729.8**
B. Selection diversity[a]	4	63.2**	299.3**	320.6**
C. Centroid distance[a]	3	44.3**	206.4**	19,532.2**
D. Number of indicators	3	0.5	0.3	1,174.7**
A × B	8	0.9	2.3	58.9**
A × C	6	1.4	1.0	2,928.7**
A × D	6	2.7*	2.2	63.6**
B × C[a]	12	6.3**	45.6**	50.6**
B × D	12	0.6	0.7	0.7
C × D	9	2.1	1.3	167.4**
A × B × C	24	1.6	1.6	10.5**
A × B × D	24	1.4	0.9	1.0
A × C × D	18	1.4	1.1	13.6**
B × C × D	36	0.7	0.5	0.8
A × B × C × D	72	0.9	0.8	1.0
Full model *F* value	239	3.0	10.8	510.6
Variance explained		1.5	5.1	71.9
Grand mean		.01	−.02	−.11

Note. Error degrees of freedom = 47760; CFA = confirmatory factor analysis with iterated maximum-likelihood estimation; EFA = exploratory factor analysis with iterated maximum-likelihood estimation. Domain communality is the height of the domain sampling sphere on the construct axis (.4, .6, or .8). Selection diversity is the maximum width of the domain sampling sphere (.1, .2, .3, .4, or .5). Centroid distance is the degree of correlational overlap between the centroids of each construct's respective indicator domains (0, .3, .5, or .7). Number of indicators varied from three to six for both constructs; the two-indicator case was not included here. The dependent variable is the difference between the obtained interfactor correlation and the true population correlation.
[a] The nature of this trend is depicted in Figures 4, 5, and 6.
$^{*}p < .01$. $^{**}p < .0001$.

recoverability of the true correlation was systematically related to characteristics of the indicators, with the remaining 98.5% of the variance simply the random sampling variability associated with our simulated selections. Generally speaking, this minimal bias was in the direction of a very slight overestimation of the true correlation (.01, or .01%). Given that we did not allow dual loadings or correlated residuals, this degree of bias (or, rather, nonbias) is quite encouraging for the confirmatory technique.

Regarding the principal axis or common-factor technique, the results showed that a little more than 5% of the variance in the recoverability of the true correlation was attributable to the influence of the four dimensions of indicator selection. Again speaking generally, the bias was in the direction of a slight underestimation of the true correlation (−0.02, or .04%; see Table 2).

In stark contrast, the raw-data estimates showed extreme bias, as expected. Here, over 70% of the resulting information about the relations between the two constructs was attributable to the four selection dimensions. Given that it does not account for measurement error, this technique, as is well known, has a pronounced tendency to underestimate the true correlation (−0.11, or 1.2%). These results highlight the fact that relying on raw-data techniques to draw inferences about the nature of the relations among constructs will be biased, not only because of measurement error but also because of the influences of the other selection dimensions (see Table 2).

Effects of selection diversity and centroid distance. The two dimensions of indicator selection that had the largest and most systematic influence on the analytic techniques were the degree of selection diversity and the degree of centroid distance (i.e., the true correlation between the construct centroids). Figures 4, 5, and 6 depict the two-way interactions between selection diversity and centroid distance for the analytic techniques. The natures of both the main effects and the interactive effects are evident in these figures.

Regarding the confirmatory technique (Figure 4), the degree of bias was uniformly small. On average, the confirmatory approach recovered the true correlation among the constructs. In no case did this technique show evidence of a negative bias, and the small degree of positive bias emerged only under conditions

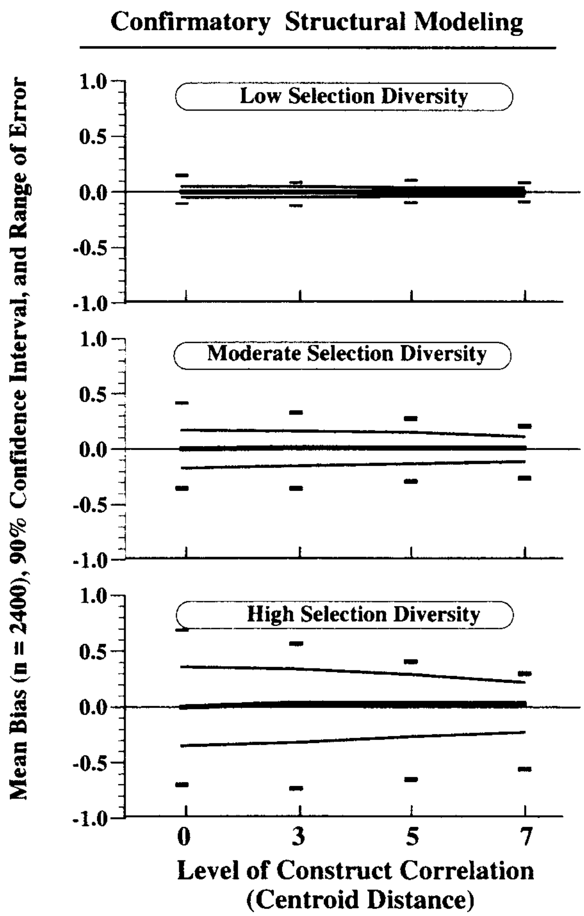

Figure 4. Mean, 90% confidence interval, and absolute range obtained from the confirmatory factor analysis of the simulated data. (Note. The plotted values are the differences between the true correlation and the obtained correlation. The observations have been collapsed across selection communality [.4, .6, and .8] and number of indicators [3, 4, 5, and 6], thus sample size for each plotted condition is 2,400. The center line is the degree of bias, the dash mark at each level of interconstruct correlation reflects the absolute range of the possible correlations, and the parallel solid lines are the 90% confidence intervals.)

of high selection diversity and low centroid distance (i.e., high correlation). Because we did not adjust our models to account for correlations among residuals or allow dual construct loadings for the indicators, this small degree of bias is particularly comforting. For the exploratory technique (Figure 5), a degree of negative bias exists in the estimates at all levels of selection diversity. The degree of bias is most pronounced at high levels of interfactor correlation (i.e., low centroid distance).[7] For the raw-data technique (Figure 6), the degree of bias was quite pronounced because, as mentioned, this approach does not account for the unreliability of the indicators.

Effects of selection communality and number of indicators. From a bias standpoint, the number of indicators per construct and the magnitude of the loadings (selection communality) had little effect on the recoverability of the true correlations. However, these two dimensions did have a pronounced influence on the efficiency (variability) of the analytic techniques in recovering the correlation between the two constructs. Fewer indicators and lower communality (both of which reflect potentially low reliability) led to quite sizable variability in the estimated correlations but did not systematically bias these estimates.

The degree of selection communality (either as a main effect or in interaction with the other three dimensions) had a pronounced influence only on the raw-data technique, given that it does not estimate the amount of reliable construct-related variance for a set of indicators. This basic feature of these techniques is well known. However, the severity of the bias that can emerge is quite startling and highlights the degree of potential error associated with conclusions and statements of generalizability using this approach.

The case of two indicators. As mentioned, we included the two-indicator case in our design because it is a relatively common situation facing researchers. As presented in Table 1, the two-indicator case led to a great deal of variability in the estimated correlation between the two constructs. More noteworthy, perhaps, is the general problem of two indicators when one attempts to analyze them. For each analytic technique except the raw-data technique, the constructs were not properly identified such that an optimal and

[7] Although the variability of the exploratory procedures was less than that of the confirmatory technique, this did not compensate for the bias differences between the techniques. In addition, the variability of the confirmatory technique is mostly due to the highly restrictive nature of the model. That is, the exploratory procedures are saturated estimates of the relations among the constructs, whereas the confirmatory model is highly restricted. If one were to examine the fit of these models (e.g., residuals, modification indexes) and allow appropriate estimates (e.g., correlated residuals among indicators that emanate from neighboring regions of multivariate space), it is likely that the variability in the estimated correlations would be considerably reduced (and perhaps the slight positive bias would be eliminated). Follow-up work is currently underway to examine the influence of these and other forms of model modification in the context of indicator selection.

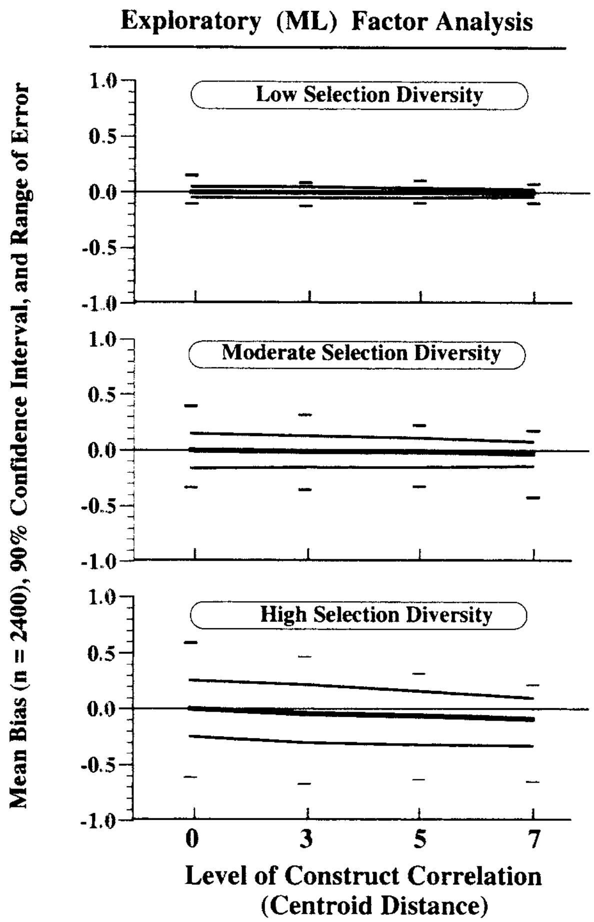

Figure 5. Mean, 90% confidence interval, and absolute range obtained from the exploratory factor analysis of the simulated data using a principal axis with maximum-likelihood (ML) estimation. (Note. The results for the principal-axis solution with least-squares estimation were nearly identical to those depicted here. See also the note to Figure 4.)

unique solution could be recovered, and the overall bias that emerged was quite pronounced. However, for the confirmatory technique, we also estimated the models with equality constraints on the loadings of the respective constructs' indicators (tau-equivalence constraints). Given that both indicators were random selections from the same domain, the assumption of tau equivalence is justified. The results of this constraint were quite encouraging for the confirmatory technique, showing a pattern of minimal bias mirroring that presented in Figure 4.

Discussion

Selecting indicators for modeling with latent variables is, and will remain, a crucial concern in experimental design. As demonstrated in our simulation, choices related to this issue strongly and interactively influence the accuracy of conclusions about relations among constructs. In particular, the four design aspects that were defined and incorporated into the simulation were all found in one way or another to influence both the representation of the constructs and the recoverability of the construct relations. In various ways, our results support and empirically reinforce a number of basic assumptions and principles that have been acknowledged in the literature for some time and highlight several points to keep in mind in designing research involving latent variables. In addition, the results provide several answers, or directions from which answers might come, to a number of prevailing questions about using multiple indicators to represent constructs. In the following, we address some of these questions in the form of general recommendations.

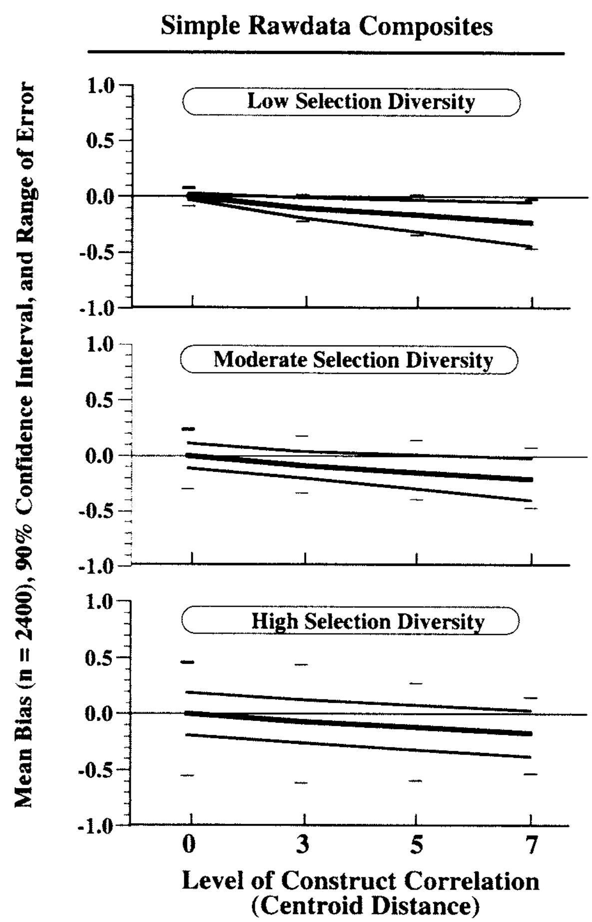

Figure 6. Mean, 90% confidence interval, and absolute range obtained from the simple raw-data estimation of the correlation. (Note. See also the note to Figure 4.)

Recommendations and Cautions

Should confirmatory or exploratory factor analysis be used for fitting the measurement model? Our data suggest that when a priori understanding of the domain allows, modeling and estimation should incorporate known information rather than rely on data-driven representations. Overall, our results demonstrated that confirmatory analyses provide less biased—in fact, nearly unbiased—estimates of the construct correlations than do exploratory analyses. Exploratory analyses approximated the lack of bias achieved with the confirmatory analysis in two cases, but they never exceeded it. The first case was under conditions of low or medium diversity, in which no bias (low diversity) or only little bias (medium diversity) was observed. The second case concerned all levels of diversity when the true construct correlation was zero. Under this condition, none of the methods produced any substantial degree of bias. In all other situations, however, the exploratory method underestimated the true construct correlation to a noticeable degree. Therefore, when one is forced to explore rather than to take a confirmatory stance, the results of the exploratory procedures must be considered with reservation and evaluated in relation to the biasing influence that the dimensions of indicator selection can have.

Do confirmatory techniques overcorrect for measurement error? In contrast to exploratory techniques, bias was basically absent with the confirmatory method, with the exception of a small positive bias under high diversity conditions coupled with high construct correlations (i.e., when diverse domains overlap considerably). On the whole, however, we found that the discrepancies in the magnitudes of construct correlations between confirmatory and exploratory techniques are due, on average, to a negative bias of the exploratory techniques, rather than to a positive bias of the confirmatory techniques. Contrary to prevailing skepticism, we found little reason to question the magnitude of a correlation between constructs based on confirmatory techniques just because it greatly exceeds the magnitude of the corresponding correlation obtained with other analytic techniques.

How many indicators are optimal? In general, using more rather than fewer indicators to define constructs produces a more efficient representation of the constructs and their interrelationships, although, on average, the number of indicators showed little effect of bias. The greater efficiency yielded through more indicators is consistent with general sampling notions. However, because the precision and accuracy of domain specifications and variable sampling do not yet approach those of sampling persons or other entities, more precise specification of content domains must be developed before the matter of sampling variables is as straightforward as that of sampling persons or other entities (see also Velicer & Fava, 1987).

Practically speaking, the optimal versus maximal principle of parsimony must be considered (see McArdle, 1994). In this regard, we do not want to imply that random sampling is the only possibility here. Random sampling is an efficient way to produce representativeness under conditions in which the sample space is well defined and the number of elements sampled is large enough to give acceptable standard errors. On the other hand, when a given domain is well enough specified that one can deliberately select (rather than sample), a small number of indicators may suffice to identify the construct precisely. In such situations, a large number of indicators is not necessary, and one can trade off variables in order to increase the number of other design elements such as persons or occasions of measurement in combinations that are optimal for one's research purposes.

Here, such deliberate selections can be informed by considering the dimensions of indicator selection that we have articulated. All else being the same, one will obtain more precise estimates of construct interrelationships when the indicators that define a construct surround its domain centroid (see Figure 1B). Again, this idea presumes that a construct's domain is well enough specified so that the notion of a domain centroid makes sense. For instance, both hierarchical (Carroll, 1993) and radex theories of intelligence (Marshalek et al., 1983) indicate that measures of reasoning ability tend to cluster closely around the centroid of the intellectual-ability domain (e.g., they load highly on Stratum III in Carroll's terminology and are located close to the center in the Marshalek et al. conception). Therefore, if one wants to obtain a sufficient estimate of individual differences in general aspects of intellectual functioning using only a few indicators, a good choice is to sample measures within the subdomain of reasoning, which is likely to triangulate on the centroid of the intelligence construct.

Under which conditions are homogeneous indicators preferred? Given the guidance of a strong theory or past research to support the contention that the sets of indicators being used to mark the constructs capture the centroids and given also that the relationship between constructs is to be estimated with measurement models specified and fitted by means of confirmatory analysis, one is better off with highly correlated indicators (i.e., high domain communality and low selection diversity). Under these conditions, such indicators are most likely to yield unbiased and efficient representations of constructs and their interrelations. With the guidance of strong theory, high selection communality and low selection specificity can be achieved with, for example, scales that are comprised of many items or numerous indicators that are aggregated into domain representative parcels (Kishton & Widaman, 1994).

Under which conditions are heterogeneous (diverse) indicators preferred? Given a theory that is not strong enough to guide the selection of specific indicators a priori but is at least capable of identifying plausible loading patterns, one is still better off using confirmatory techniques to fix the measurement model. In this case, however, the selection of indicators should generally span the domain rather than be highly targeted (cf. Humphreys, 1962). Often, the temptation is to maximize homogeneity, which can lead to representations of constructs that are sharply defined (i.e., highly intercorrelated indicators) but may be off-center with regard to the domain centroid.

For example, consider selection planes of moderate diversity. Attempting to reduce the diversity among indicators by selecting only those indicators that contribute to high estimates of internal consistency may result in a construct that reflects not only true selection communality but also a considerable amount of nonconstruct specificity (e.g., common method variance, response biases, social desirability). The allure of a bloated-specific construct is nefarious because the likelihood is that the sampled centroid among the indicators, although well marked, would be biased (i.e., when "good" indicators are bad). This observation may seem contradictory to those persuaded of the unqualified virtues of homogeneity (i.e., high correlations among the indicators of a latent variable). However, choosing a set of indicators mostly on the basis of high intercorrelations runs a significant risk of both missing the centroid altogether and wreaking havoc on the estimates of construct relations (i.e., when "good" indicators are bad). Thus, contrary to common belief, high indicator intercorrelations are not always better than low indicator intercorrelations. This situation is typical of broad domains whose centroid, for whatever reason (insufficient data or theory, lack of convergence of data and theory, diffuse structure of the domain itself), is not well identified.

Can indicators with low internal consistency still be valid? Even a set of indicators of "poor" psychometric quality (e.g., low reliabilities, little common variance) can produce accurate estimates of the relationships among constructs provided that they (a) are spread out across the construct domain sufficiently to capture the centroid, (b) yield sufficient variability on the construct, and, again, (c) are analyzed by confirmatory analysis. Indicators of constructs such as socioeconomic status or aspects of attachment, for example, often are faulted for their low interrelations. Arguably, constructs such as these might better be conceived of as emergent constructs, and, if so, our discussions would not apply. However, if such constructs are presumed to give rise to effect indicators, our simulation suggests that a confirmatory representation of such indicators not only triangulates on the construct's centroid but also, on average, accurately corrects for the construct's low measurement quality (disattenuates) and yields unbiased estimates of its relations with other constructs. In other words, constructs can be represented validly even though estimates of reliability suggest otherwise (i.e., when "bad" indicators are good).

What should be done with just two indicators? The increasing sensitivity of researchers to the problems associated with two indicators has brought forward the question of how to appropriately handle such cases in actual practice. Having only two indicators to identify a construct has been recognized as problematic for some time (e.g., Harman, 1967). In the context of our geometric framework, the fundamental flaw of having only two indicators is evident. If one locates two indicators on the periphery of a selection plane (e.g., as in Figure 1B), the line segment that joins them contains the sample centroid; however, a centroid's location on that line segment is not determinable without additional information. Even the line segment between two diametrically opposed indicators, although containing the true centroid, would not have a uniquely defined sample centroid. Placing theoretically meaningful constraints on the defining equations of two-indicator constructs, however, uniquely locates a centroid for that sampling of indicators. For example, we suggest, on the basis of our simulation, that when two indicators of a construct are theoretically equivalent selections from the domain of possible indicators, placing equality constraints on the respective loadings is theoretically defensible (even if the empirical data are not in strong agreement) and, on average, leads to accurate recovery of the true construct centroids.

What can be done with nonoptimal indicators? When one is confronted with a less than optimal set of manifest variables for marking the constructs of interest, any additional information that can be brought to bear in estimating the construct interrelationships should be used. Specifying the representation of the constructs as precisely as possible can help to compensate for ill-chosen indicators. For example, the sample centroid can be weighted toward indicators of higher validity if the validities are known. More generally, other weighting schemes can be used to estimate sample centroids that in turn will produce more valid estimates of the relationships among constructs. For instance, in the case of two indicators, if one of them is a more theoretically important marker of a construct than the other, placing equality constraints on the loadings of the indicators is not appropriate. However, the lack of identifiability and the possibility that the non-construct-related variance from the weak indicator may "sneak" its way into construct space suggest that one should still place theoretically meaningful constraints on the defining equations of the construct. Here, the central indicator could be constrained by fixing its loading and its residual error to theoretically defensible and meaningful values (e.g., the loading could be constrained to equal the square root of a communality estimate, and the error could be constrained to 1 − communality).

How can one better select indicators? On the basis of our framework and the simulation results, as well as on the principles of convergent and discriminant validity, we can suggest an empirically aided heuristic for selecting optimal indicators of a construct. First, on the basis of theory and available research, specify the expected location of a construct centroid relative to a set of key marker constructs. For example, a construct under development might be expected to be moderately positively correlated with established Construct X, moderately negatively correlated with established Construct Y, and uncorrelated with established Construct Z. Second, identify a broad set of candidate indicators of the construct and use them to estimate a strict confirmatory factor analysis of the four constructs (i.e., with no unnecessary estimates or restrictions). On the basis of our simulation, the correlational structure among the constructs should be an unbiased and quite efficient (if sufficient indicators are used) representation of the constructs' interrelations. With this construct space as a guide, then select a subset of the indicators of the construct under development that, when placed in a strict confirmatory factor analysis, yields the same correlational structure among the constructs. We emphasize that, with this approach, the optimal subset of indicators of the construct, again, may or may not be the same subset of indicators that would yield the highest internal consistencies.

Is more always better? In the face of limited or nonexistent theory and given no other information, two recommendations seem warranted to us: More is better, and breadth is better than homogeneity. However, the idea that more is better is not necessarily a take-home conclusion that can be deduced from our framework. One important goal in programmatic research is to find an optimal and unbiased set of indicators that allow accurate and efficient representations of constructs. As was shown in Table 1, six indicators under high diversity conditions are as unbiased and efficient as two indicators under moderate diversity conditions. Therefore, the more-is-better notion must be considered in conjunction with the expected diversity of the domain. Under high diversity, more items or indicators are certainly better if prior research or strong theory is not available to assist in selecting an optimal set of indicators. Under low to moderate diversity, parsimony and optimality can make for good representations. In addition, we mention that preanalysis procedures such as parceling techniques offer a positive feature in that they reduce the diversity of the indicators. Given that some recent works have emphasized that three indicators per construct in confirmatory analyses is an optimal number, parceling techniques can bring the number of indicators down to this optimum while reducing the diversity to a level at which the construct representation will most likely be accurate (see, e.g., Kishton & Widaman, 1994; Marsh, Hau, Balla, & Grayson, 1998).

Conclusions

As should be apparent throughout our discussion, the strength of our recommendations rests considerably on the availability of a strong theory concerning the nature of the construct domain and the dispersion of indicators through it. In our view, theory provides the paramount source of guidance for picking a limited number of indicators to represent a construct. Whether theory provides actual indicators or merely defines a domain that can be representatively sampled is a secondary question. In either case, because theory is never complete, we must recognize the importance of the (hopefully) spiraling interplay between theory and empirical research whereby both are improved gradually (Cattell, 1966a).

In our view, measurement will be considerably improved if researchers consider the centroid of a construct that they desire to measure and speculate on its place among other constructs rather than getting caught up in the intricacies of the correlations and consistencies among the measured items, behaviors, or responses. Shifting attention from the trees to the forest should free the embedded figure from its frame and allow for better representations of constructs. In other words, focusing attention on the centroid of a construct and then considering the various indicators that one could select to represent the construct will, in our view, yield better choices than would, for example, selecting a set of indicators by convenience or on the basis of high internal consistency estimates and then presuming that an analysis technique will recover the true structure. Granted, all analysis techniques attempt to optimize the implied structure inherent in a set of manifest indicators. However, their efficacy will be only as good as the measurements provided. If the indicators are not selected well, the recovered structure will be misleading and biased.

Regarding limitations and future directions, the implications of our simulation apply most directly to variables that are multivariate normal and continuous. We do not know, for example, how far the implications can be extended to nonnormal data or to ordinal variables. A second issue for further work has to do with the possibility of estimating correlated residuals and dual construct loadings in structural-equations frameworks. To what degree would additional estimates such as these contribute to or diminish the already low degree of bias with such techniques as well as increasing the efficiency of estimation? A third issue has to do with representing mean-level, or intercept, information in structural-equations models (i.e., MACS models; Little, 1997). Do these four selection dimensions have a differential influence on estimates of the mean-level information?

These caveats notwithstanding, the results of our simulation indicate that strong theory coupled with a confirmatory approach will, on average, succeed in recovering the true multivariate relations among indicators and constructs, even under difficult conditions. We emphasize, however, that, underlying the specifics of the selection framework and Monte Carlo simulation, our more general aim has been to reinforce the need for continuing efforts to develop and elaborate the definition of constructs and thereby to improve the measurement devices that are such an important aspect of our science.

References

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society, 4,* 106–110.

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 219–251). New York: Academic Press.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods.* Hillsdale, NJ: Erlbaum.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419–456.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48,* 386–398.

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

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

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

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1–76). Chicago: Rand McNally.

Carroll, J. B. (1993). *Human cognitive abilities.* Cambridge, England: Cambridge University Press.

Cattell, R. B. (1952). *Factor analysis.* New York: Harper.

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. (1966a). The data box: Its ordering of total resources in terms of possible relations systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 12–28). Chicago: Rand McNally.

Cattell, R. B. (1966b). Multivariate behavioral research and the integrative challenge. *Multivariate Behavioral Research, 1,* 4–23.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement, 24,* 3–30.

Cerny, B. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlations matrices. *Multivariate Behavioral Research, 12,* 43–47.

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

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Latent variables in structural equations' causal models. *Applied Psychological Measurement, 14,* 183–196.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dumenci, L., & Windle, M. A. (1996). A latent trait-state model of adolescent depression using the Center for Epidemiologic Studies-Depression Scale. *Multivariate Behavioral Research, 31,* 313–330.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Gorsuch, R. L. (1988). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Harman, H. H. (1967). *Modern factor analysis.* Chicago: University of Chicago Press.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401–415.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, mark IV. *Educational and Psychological Measurement, 34,* 111–117.

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.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, 62,* 28–30.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworths.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 343–357.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1–15.

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

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

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

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.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7,* 107–127.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions, 5,* 11–18.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 234–251.

McArdle, J. J., & Woodcock, R. W. (1996). *Expanding test–retest designs to include developmental time-lag components.* Unpublished manuscript, University of Virginia.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

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

Messick, S. (1982). Test validity and the ethics of assessment. *Diagnostica, 28,* 1–25.

Mulaik, S. A. (1988). Confirmatory factor analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 259–288). New York: Plenum.

Nesselroade, J. R., & Jones, C. J. (1991). Multi-modal selection effects in the study of adult development: A perspective on multivariate, replicated, single-subject, repeated measures. *Experimental Aging Research, 11,* 21–27.

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

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

Ozer, D. J. (1986). *Consistency in personality: A methodological framework.* Berlin: Springer-Verlag.

Pearson, K. (1903). Mathematical contributions to the theory of evolution—XI: On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London (Series A), 200,* 1–66.

Poon, L. W., Krauss, I. K., & Bowles, N. L. (1984). On subject selection in cognitive aging research. *Experimental Aging Research, 10,* 43–49.

SAS Institute. (1990). *SAS/STAT user's guide* (Vol. 1, 4th ed.). Cary, NC: Author.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8,* 79–98.

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.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 22,* 193–209.

Widaman, K. F. (1992). Multitrait–multimethod models in aging research. *Experimental Aging Research, 18,* 185–201.

Received July 11, 1997
Revision received June 23, 1998
Accepted September 4, 1998 ■