MULTIVARIATE BEHAVIORAL RESEARCH, *42*(2), 349–386

Applying the Bootstrap to Taxometric Analysis: Generating Empirical Sampling Distributions to Help Interpret Results

John Ruscio
The College of New Jersey

Ayelet Meron Ruscio
University of Pennsylvania

Mati Meron
University of Chicago

Meehl's taxometric method was developed to distinguish categorical and continuous constructs. However, taxometric output can be difficult to interpret because expected results for realistic data conditions and differing procedural implementations have not been derived analytically or studied through rigorous simulations. By applying bootstrap methodology, one can generate empirical sampling distributions of taxometric results using data–based estimates of relevant population parameters. We present iterative algorithms for creating bootstrap samples of taxonic and dimensional comparison data that reproduce important features of the research data with good precision and negligible bias. In a series of studies, we demonstrate the utility of these comparison data as an interpretive aid in taxometric research. Strengths and limitations of the approach are discussed along with directions for future research.

Meehl's (1995) taxometric method is designed to assist researchers in determining whether the latent structure of a variable is categorical or continuous. Each

Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628. E-mail: ruscio@tcnj.edu

taxometric procedure within the method yields one or more curves whose shapes are inspected to inform a structural inference. For example, several procedures yield peaked curves in the presence of latent groups and nonpeaked curves when latent dimensions, rather than groups, underlie the observed scores. The developers of the method (e.g., Grove, 2004; Meehl, 1995; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998) have described the curve shapes that one would expect to observe for categorical vs. continuous latent variables. Rather than relying on tests of statistical significance, researchers inspect taxometric curves for these characteristic shapes, with confidence in a categorical or dimensional inference increasing with greater consistency of results across mathematically diverse analytic procedures.

Monte Carlo studies of the validity of inferences based on taxometric analyses are scarce, perhaps in part because taxometric curves are more challenging to interpret mechanically than the strictly quantitative output of many other procedures. Whereas numerous publications have presented prototypical curve shapes for categorical and continuous data, only two Monte Carlo studies have presented and evaluated curve shapes across systematically varying conditions (Meehl & Yonce, 1994, 1996). Unfortunately, the generalizability of these two studies was constrained by their restriction to highly idealized data conditions that did not span the breadth or complexity of conditions likely to be encountered by researchers. In addition, because each taxometric procedure was implemented in only one way, these studies were unable to determine the influence on curve shapes of the many choices that must be made to translate a taxometric procedure, presented in general terms, into a concrete algorithm for data analysis (see Grove, 2004, footnote 6). As a result, researchers attempting to use the results of these studies to guide the interpretation of their own taxometric curves must extrapolate well beyond the conditions studied into uncharted territory. At present, little is known about the appearance of taxometric curves—and how best to interpret these curves—under realistic research conditions.

We propose that researchers take advantage of computing power to obtain comparative results that are not available through analytic derivations or in extensive simulation studies. In particular, an approach known as the *bootstrap* (Efron, 1979) is well-suited to the type of problem facing researchers who wish to use the taxometric method under unique data conditions, with a procedural implementation that has not been directly studied.[1] The bootstrap involves

[1]The use of the term "bootstrap" throughout the present article refers to the resampling strategy introduced by Efron (1979) to study the performance of a statistical procedure using data-based estimates of the relevant population parameters. Cronbach and Meehl (1955) used the term to describe the process by which one could develop measures more valid than the criterion variables that were originally used to validate them, and Meehl (1995) has described his taxometric method in a similar way. Thus, the approach to taxometric analysis that we suggest involves both types of bootstrapping.

data-based simulation, which allows one to study the performance of an analytic procedure—implemented in any way that one chooses—using data-based estimates of the relevant population parameters. Efron and Tibshirani (1993, p. 395) explain that "the bootstrap is used not to learn about the general properties of a statistical procedure, as in most statistical simulations, but rather to assess its properties for *the data at hand*" (emphasis in original). Since its introduction by Efron (1979), investigators have used the bootstrap to estimate statistics that could not be derived analytically or have not been studied adequately using Monte Carlo methods (Efron & Tibshirani, 1993). In recent years, the bootstrap has been applied to multivariate procedures such as item-response theory (e.g., Muthén, 2001) and latent variable mixture modeling (e.g., Stone, 2000).

Because Monte Carlo studies have not established the expected curve shapes for categorical and continuous data across realistic research conditions, there is considerable potential in applying the bootstrap to Meehl's taxometric method. The bootstrap can be used to generate empirical sampling distributions of taxometric curves that take into consideration both data-based estimates of the population parameters and a specific procedural implementation. One's research results can be compared to these distributions to help judge whether the obtained curves are more consistent with categorical or continuous latent structure. While a number of researchers, beginning with Gangestad and Snyder (1985), have recognized the value of simulated comparison data for taxometric studies, we are not aware of suggestions that researchers simulate comparison data representing both of the competing structural models that the taxometric method aims to distinguish—a central tenet of our approach.

In what follows, we briefly review Meehl's taxometric method, describe one illustrative taxometric procedure, and discuss how the bootstrap can be applied to taxometrics. After presenting an algorithm for generating bootstrap samples of comparison data, we report three simulation studies, one evaluating the precision and bias with which this algorithm reproduces important data characteristics, and two examining the value of using bootstrap data to inform the interpretation of taxometric results.

MEEHL'S TAXOMETRIC METHOD

Behavioral scientists have long struggled with the classification problem—the question of where qualitative distinctions can be drawn between latent classes (types, categories, or taxa) and where quantitative differentiation exists along latent factors (continua, traits, or dimensions). This important problem presents some daunting challenges that are well captured by the deceptively simple task addressed by Meehl's taxometric method: distinguishing a *taxonic* structural model comprising two latent classes (the *taxon* and *complement*, within which

there may also be variation along one or more latent factors) from a *dimensional* structural model comprising one or more latent factors. Both structures can give rise to similar observed variances and covariances of measured variables (or *indicators*), rendering simple inspection of distributions or correlations ineffectual for drawing valid structural inferences (Murphy, 1964). Whereas the expected variance of an indicator of a dimensional construct—as well as its covariance with other indicators—depends on its pattern of shared and unique factor loadings, the expected variance of an indicator of two latent classes is:

$$s_x^2 = Ps_1^2 + Qs_2^2 + PQD_x^2, \tag{1}$$

where s_x^2 is the variance of an indicator x in the total (mixed-group) sample, s_1^2 and s_2^2 are the variances within the two classes, P is the base rate of the taxon, $Q = 1 - P$, and D_x^2 represents the square of the (unstandardized) mean difference between the classes on indicator x (Waller & Meehl, 1998). Similarly, the expected covariance of two such indicators is given by the General Covariance Mixture Theorem (GCMT; Waller & Meehl, 1998):

$$\text{cov}(xy) = P\,\text{cov}_1(xy) + Q\,\text{cov}_2(xy) + PQD_xD_y \tag{2}$$

Thus, one can postulate either a taxonic or a dimensional model to explain the variances and covariances observed for a given set of indicators. In fact, Bartholomew (1987) showed that models with m latent classes or $m - 1$ latent dimensions can reproduce any observed variance-covariance matrix equally well. Meehl's taxometric method is used to determine whether the relations among a set of indicators are more consistent with a taxonic or a dimensional model of latent structure.

The taxonic model is embodied in Equations (1) and (2), which represent the variance and covariance of indicators as the sum of within-group sources (the first and second terms on the right of these equations) and between-group sources (the third term). Within the taxonic model, the pertinent model parameters are conventionally referred to as the *taxon base rate* (P; with complement base rate $Q = 1 - P$), *indicator validities* (D_x and D_y, the mean differences—or separation—between taxon and complement classes on indicators x and y), and *nuisance covariance* (the covariance between indicators x and y within each class). In practice, indicator validity and nuisance covariance are often estimated and expressed in standardized units so that they can be evaluated independently of sample-specific measures. Indicator validity is estimated using the metric of Cohen's d, whereas nuisance covariance is estimated using the Pearson product-moment correlation.

The dimensional model corresponds to the common factor model in which an indicator variance-covariance matrix is modeled using the shared loadings of k indicators on n latent factors ($1 \leq n \leq k$). Taxon base rate, indicator validity, and

nuisance covariance do not denote parameters in the dimensional model because these require more than one latent class. However, for any configuration of taxon base rate, indicator validities, and nuisance covariance among a set of indicators, the indicator correlation matrix can be reproduced using the common factor model (Bartholomew, 1987). In this way, nonoverlapping sets of parameters involved in two structural models can be calibrated to yield similar indicator correlation matrices (see Meehl & Yonce, 1994, p. 1146).

Each analytic procedure within the taxometric method examines the relations among indicators to provide clues about latent structure. For example, the MAXEIG (MAXimum EIGenvalue; Waller & Meehl, 1998) procedure examines the association among two or more *output* indicators within subsamples ordered along an *input* indicator. First, cases are sorted by their scores on the input indicator and grouped into ordered subsamples through a series of overlapping windows. For instance, in a sample of $N = 1{,}000$ cases, using 91 windows that overlap 90% with one another yields subsamples of (sorted) cases numbered 1–100, 11–110, 21–120, . . . , 901–1000. Second, within each subsample, a modified covariance matrix is constructed using the output indicators. This matrix consists of a variance-covariance matrix whose diagonal elements are replaced by zeros to remove variances and leave only covariances. Third, the first (largest) eigenvalue of this modified covariance matrix is calculated within each subsample. Finally, a graph is constructed by plotting subsample means on the input indicator along the abscissa and subsample eigenvalues along the ordinate. MAXEIG curves constructed in this way are expected to take on different shapes for taxonic and dimensional data (Waller & Meehl, 1998): Eigenvalues are expected to remain constant across subsamples for dimensional data, but to rise to a peak for taxonic data. The emergence of a taxonic peak stems from the fact that, as shown in the GCMT, the mixture of groups increases the covariance between any two valid indicators of taxonic structure. Thus, subsamples in which groups are mixed in roughly equal proportions should yield the largest indicator covariances, and therefore the largest eigenvalues. As subsamples become increasingly homogenous (i.e., all or most cases are members of one group), indicator covariances and eigenvalues should be lower. Thus, a peaked MAXEIG curve suggests the mixture of groups, whereas a nonpeaked MAXEIG curve suggests the absence of groups.

All taxometric procedures are expected to produce differently shaped curves for taxonic and dimensional data. The proper interpretation of obtained curves is the central challenge of a taxometric investigation. Although the prototypical curve shapes for taxonic or dimensional data are well established, the complexities of actual research data often yield curves that are more difficult to interpret. Not only can these curves look quite different from those expected for idealized data, they can be misleading under certain conditions. For example, when indicators of a taxonic construct do not separate groups with sufficient validity,

the resulting curves can look just like the prototypes for dimensional structure (e.g., a flat MAXEIG curve). Or, when indicators of a dimensional construct are positively skewed, the resulting curves can be mistakenly interpreted as taxonic (e.g., a MAXEIG curve that rises to an apparent peak at its right end; J. Ruscio, Ruscio, & Keane, 2004).

For these reasons, it may be very useful to researchers to have a sense for how taxonic and dimensional curves are likely to look under the unique conditions of their particular study. Our application of the bootstrap to taxometric analysis is intended to serve exactly this purpose. It provides a benchmark for comparison that is tailored to the specific data and analytic conditions faced by the researcher, and consequently may be less likely to result in a mistaken structural inference than relying solely on the idealized curves in Monte Carlo studies. Our proposed approach not only may help in determining whether research results are more consistent with taxonic vs. dimensional structure, but may show a particular test to be uninformative (i.e., unable to distinguish the curves for comparison data known to be taxonic and dimensional) or ambiguous (i.e., finding the obtained curve to be equally consistent with taxonic and dimensional curves), preventing inferences from being drawn under these conditions.

TECHNIQUES FOR BOOTSTRAPPING COMPARISON DATA

In order to obtain empirical sampling distributions of taxometric curves, which represent the expected curve shapes for competing structural models under the conditions of a particular study—one must first generate, then analyze bootstrap samples of taxonic and dimensional comparison data. Using the bootstrap to generate univariate distributions is straightforward: From a sample of N scores on a single variable, one randomly resamples (with replacement) N values (Efron & Tibshirani, 1993). This technique treats the sample distribution as an unbiased estimate of the population distribution, from which one then draws B random samples, each of size N.

Bootstrapping multivariate distributions, on the other hand, requires careful consideration of the technique by which the dependence among variables is reproduced. This can be illustrated by the application of the bootstrap to the more familiar task of estimating confidence intervals (CIs). For example, to empirically estimate a 95% CI for the correlation between variables x and y, one can (1) randomly resample (with replacement) N pairs of x-y values to generate each of B bootstrap samples of bivariate data, (2) calculate the correlation in each of the bootstrap samples, and (3) select the 2.5^{th} and 97.5^{th} percentiles within the sorted distribution of these B correlations as the lower and upper limits of the CI (see Efron & Tibshirani, 1993, for details on this and other ways

to generate bootstrap CIs). This technique treats the observed distribution of bivariate scores as an unbiased estimate of the population distribution and draws bivariate samples accordingly. A sufficiently large value of B (e.g., $\geq$1,000) is required to obtain stable estimates of the tails of the empirical sampling distribution. The empirically derived CI (as opposed to one that is derived analytically) makes no assumptions about the univariate or bivariate distributions of x and y; instead, it reproduces the distributions observed in the research data by resampling from an unbiased estimate of the population distribution. If one's objective is, instead, to test the statistical significance of a correlation, one can generate bootstrap samples by randomly resampling (with replacement) x and y values independently, rather than in pairs (Lee & Rodgers, 1998). Doing so models the null hypothesis of $\rho = 0$ while reproducing the observed univariate distributions, again resampling from unbiased estimates of population distributions. Lee and Rodgers (1998) found that, across varying sample sizes, population correlations, and population distributions, hypothesis tests using univariate rather than bivariate bootstrap samples achieved a better combination of statistical power and control over Type I error rates. Based on these findings, Lee and Rodgers argued that whereas bivariate resampling is consistent with the logic of CI construction and appropriate for that purpose, superior tests of the null hypotheses could be obtained using univariate resampling. For present purposes, the important point is that the choice of a bootstrap resampling strategy should be guided by one's research goals.

To apply the bootstrap to taxometrics, one must generate bootstrap samples of comparison data in ways that represent the latent structures that the taxometric method is designed to distinguish: taxonic (categorical) and dimensional (continuous). Natural choices are to use the common factor model to represent dimensional structure and the GCMT to represent taxonic structure.

Generating a bootstrap sample of dimensional data (BSD) represents the most fundamental challenge. If this can be achieved, then repeated application of the algorithm would allow one to reproduce data within two separate groups, and hence in a full sample of taxonic data. Although it is relatively simple to reproduce either indicator distributions or correlations, reproducing both is more complex. The common factor model can be used to generate normal, continuously distributed indicators correlated at desired levels. However, researchers seldom measure variables along truly continuous scales. Haslam and Kim (2002) found that that the most common source of data for taxometric studies was self-report instruments. The binary or Likert-type rating scales of these instruments yield ordered categorical variables, although the number of ordered categories may be large when items are aggregated to form composite indicators for taxometric analyses. As a result, when the common factor model is used to reproduce indicator correlations, cutoff scores must be applied to transform the continuous

indicator distributions into ordered categories. This, in turn, alters the indicator correlations, usually attenuating them.

Waller, Underhill, and Kaiser (1999) developed a sophisticated algorithm that allows researchers to draw samples from populations defined by many potentially important parameters. Their "Monte" program is intended for use in Monte Carlo studies of classification procedures, although it can also be used to generate BSDs when indicators are continuously distributed. However, the version of this algorithm that can be used to reproduce ordered categorical distributions applies thresholds to indicator distributions *after* reproducing indicator correlations, resulting in correlations that usually will be attenuated. Thus, the Monte program is not intended for the investigator who wishes to reproduce discrete distributions observed in a sample of research data (N. G. Waller, personal communication).

Our solution to the apparent chicken-and-egg problem of reproducing distributions and correlations is to use an iterative technique that combines the common factor model with univariate bootstrap indicator distributions. This technique first reproduces the desired indicator correlation matrix by applying factor loadings to normally distributed, continuous variables. It then substitutes distributions bootstrapped from the observed indicator distributions. Because the substituted bootstrap distributions will usually be non-normal (Micceri, 1989) and discrete, imposing them on the variables will alter the correlations that have been reproduced. The technique therefore iteratively updates the target correlation matrix and calculates new factor loadings, retaining the data set that best reproduces the observed indicator correlations when the bootstrap indicator distributions are imposed. The following is a step-by-step description of our algorithm for generating a BSD in this iterative manner, with variable names drawn from the "DimSample" program code presented in the Appendix:

1. Read and store the sample size N, number of indicators k, and indicator correlation matrix *Target.Corr*.
2. Generate a bootstrap distribution of scores for each indicator by resampling, with replacement, N scores from each observed univariate distribution. Store these bootstrap indicator distributions as $Freq.Dist_1 \ldots Freq.Dist_k$.
3. Store a copy of *Target.Corr* as *Desired.Corr*, which will be updated as the procedure iterates.
4. Perform a factor analysis of the target data[2] and record the number of factors with eigenvalues > 1 as *Factors*. This step determines the number

[2]Throughout this article, we use the term "research data" to refer to an actual sample of data collected in a taxometric investigation. In contrast, the term "target data" refers either to artificially generated data in our simulation studies or to the data whose parameters are estimated and reproduced using our simulation algorithms. "Bootstrap data" is different in that this refers to the output of our simulation algorithms.

of latent factors that will be used to reproduce the indicator correlation matrix, and the liberal minimum eigenvalue criterion is used so that potentially useful latent factors are not missed. Even when spurious factors are identified by this liberal criterion, loadings on these would be expected to influence only slightly the reproduction of correlations.

5. Create *Factors* vectors of N random unit normal values that will be partly shared by all indicators ($Shared.Comp_1 \ldots Shared.Comp_{Factors}$). All of our "random unit normal values" were drawn from a population with $\mu = 0$, $\sigma = 1$ using the "rnorm" function in R (Wichura, 1988).
6. Create k additional vectors of N random unit normal values that will contribute the unique, or error, component for each indicator ($Unique.Comp_1 \ldots Unique.Comp_k$).
7. Determine the weights for the shared and unique components of each indicator that will reproduce *Desired.Corr* through a factor analysis of *Desired.Corr* using *Factors* latent factors. The number of shared loadings to be saved will be $k \times Factors$. For each indicator i, $Shared.Load_{i,1} \ldots Shared.Load_{i,Factors}$ will be used to weight the shared components in a subsequent step, with the weights for each *Unique.Comp* calculated as:

$$Unique.Load_i = \sqrt{1 - \sum_{j=1}^{Factors} Shared.Load_{i,j}^2} \tag{3}$$

8. Calculate each simulated indicator ($i = 1$ to k) as a new vector of N values by weighting the shared and unique components for each by the factor loadings calculated in step #7.
9. To reproduce the indicators' estimated population distributions, replace each simulated indicator's score distribution with its bootstrap distribution. This is done by sorting the simulated data set by each indicator in turn and replacing that indicator with its corresponding sorted vector from the bootstrap distributions $Freq.Dist_1 \ldots Freq.Dist_k$. This key step preserves the rank-ordering of cases on each indicator but replaces the artificially normal, continuous simulated distribution with a bootstrap sample from the non-normal, discrete distribution observed in the research data that was used as the best estimate of the population distribution.
10. Calculate the correlation matrix *Reproduced.Corr* in the bootstrap sample, then compute a matrix of residual correlations: $Residual.Corr = Target.Corr - Reproduced.Corr$. Particularly in early iterations of the procedure, these residuals may be substantial because the potentially discrete, nonnormal indicator distributions were imposed after correlations were reproduced using the common factor model with continuous, normal data.

11. Check whether, after substituting the bootstrap distributions, the current iteration achieves the best correlational reproduction so far by calculating the root mean square residual (RMSR) correlation:

$$\text{RMSR} = \sqrt{\frac{\sum r_{residual}^2}{n_{residual}}}, \tag{4}$$

where each $r_{residual}$ represents a value below the diagonal in *Residual.Corr* and $n_{residual}$ represents the number of these values. If the RMSR correlation is the lowest so far (as on the first iteration it automatically is), store a copy of *Desired.Corr* as *Best.Corr*, store the current RMSR correlation as *Best.RMSR*, and begin (or reset) a counter j at 0; this is used to allow a finite number of additional iterations to attempt to achieve a better correlational reproduction. Empirical results revealed rapidly diminishing returns beyond 5 additional iterations, so that is the default value; users can specify an alternative value of *Trials*.

12. If the current iteration has not achieved the best correlational reproduction thus far, increment the counter ($j = j + 1$). If this marks the 5th iteration (or an alternative value specified by the user), proceed to step #13; otherwise, create a new *Desired.Corr* by adding a fraction of *Residual.Corr* to *Best.Corr*, and return to step #7. The fraction to be added is calculated as $Multiplier/2^j$, where *Multiplier* is a user-specified step-size multiplier (by default, this is set to 1.00) and j is the counter. Thus, our algorithm allows for 5 iterations to attempt to improve the reproduction of correlations. On the first try, the residual correlations are added in their entirety to *Best.Corr* (i.e., the step size multiplier $= 1.00/2^0 = 1.00$, because j is reset to 0 when the RMSR correlation is the lowest so far). If this first try fails to improve the reproduction of correlations, a second try is made using a smaller fraction of *Residual.Corr* (i.e., the step size multiplier $= 1.00/2^1 = .50$). This repeats, if necessary, for a third, fourth, and fifth try (step size multipliers = .25, .125, and .0625, respectively). If even the smallest step from *Best.Corr* fails to yield improvement, the iterative routine terminates. If any of these trials improve correlational reproduction, the counter is reset to allow up to five trials starting from a new correlation matrix.
13. Construct indicators using the factor loadings that generated correlations which, when altered by substituting the bootstrap distributions, best reproduced *Target.Corr*. These loadings are derived from a factor analysis of *Best.Corr*, which was stored for this purpose.
14. Report the RMSR correlation (stored as *Best.RMSR*) and return the bootstrap sample of dimensional comparison data.

In sum, this algorithm implements the common factor model with bootstrapped indicator distributions, achieving the reproduction of indicator correlations in an iterative manner. Although early steps in the algorithm generate and work with multivariate normal data, step #9 substitutes indicator distributions that are bootstrapped from the target data. The algorithm evaluates the extent to which substituting the bootstrap distributions has affected the reproduction of correlations, then iteratively improves this reproduction. Factors such as sample size, indicator variability (i.e., continuous vs. ordered categorical response scales), and model misspecification (e.g., incorrect number of latent factors) will determine the accuracy with which indicator correlations can be reproduced.

As outlined above, our iterative algorithm generates a BSD on the basis of a dimensional structural model. This algorithm is easily extended to generate a bootstrap sample of taxonic comparison data (BST). The GCMT involves the mixture of two groups with no shared members, and the model includes terms representing indicator covariance within one or both groups (nuisance covariance) as well as mean differences across groups (indicator validity). A BST is, therefore, generated by (1) breaking a research data set into subsamples representing two groups (taxon and complement), (2) reproducing indicator distributions and correlations separately within each subsample following the steps listed above, and (3) merging these to reconstruct the full sample. The resulting BST will match the sizes of the full sample and each group, and reproduce (a) the indicator distributions within each group (and thus in the full sample), (b) the validity of each indicator (through distributional reproduction within and between groups), and (c) the within-group and full-sample correlation matrices. The "TaxSample" program code in the Appendix shows how repeated application of the DimSample routine can yield a BST.

PROVIDING A CRITERION VARIABLE TO GENERATE A BOOTSTRAP SAMPLE OF TAXONIC COMPARISON DATA

To generate taxonic comparison data using *any* algorithm, one must possess at least a rudimentary sense of the putative class membership of cases in a data set. For example, when using the Waller et al. (1999) Monte program, users must specify the taxon base rate, each indicator's validity, the within-group indicator distributions, and the within-group indicator correlation matrices. Estimating these values for the data at hand requires the classification of cases into groups. Similarly, our iterative technique requires researchers to assign all cases to groups so that indicator distributions and correlations can be reproduced within each group.

This requirement raises some important issues. For example, it suggests that BSTs may be less useful in purely exploratory research in which there are no defensible estimates of (at minimum) the taxon base rate. However, even if there are multiple plausible taxon base rates, one can generate BSTs using case classifications that span the full range of plausible base rates to examine robustness to misspecifications of this rate.

It may be argued that because any assignment of cases to groups will be fallible, the reproduction of indicator distributions and correlations within groups represents an illusory precision. While it is true that no available classification will be infallible, this does not necessarily undermine the utility of the bootstrap. Useful classifications can be constructed in a number of ways. For instance, a knowledgeable investigator should be able to provide a plausible range of base rates for the putative taxon, and cases can be classified using these base rates alone. Although one cannot know the precise value of a taxon base rate in advance, reasonable upper and lower limits can be estimated from prior theory or research (e.g., epidemiological or other investigations), examination of frequency distributions (e.g., apparently multimodal distributions may provide hints about the relative sizes of putative groups), conventional cutting scores on widely-used measures (e.g., the proportion of individuals in the sample with a T score above 70 on an MMPI scale), a fallible external criterion (e.g., the rate of diagnosis in the sample), an educated guess, or a combination of these approaches. For a given base rate estimate P, cases can be classified by assigning a proportion P of cases with the highest indicator total scores to the putative taxon and the remainder of cases to the complement class. An alternative empirically-based technique would be to use an analytic procedure (e.g., cluster analysis) to obtain a preliminary classification of cases.

Some bias may be inevitable when cases are fallibly assigned to groups. For example, assignments based on the base-rate classification method outlined above will almost certainly produce groups that are overly homogeneous with respect to their scores on the indicators, as it is unlikely that all $P \times N$ individuals with the highest indicator total scores actually belong to the taxon or that all $Q \times N$ individuals with the lowest indicator total scores actually belong to the complement. As a result, one would expect within-group correlations (nuisance covariance) to be too low, and between-group separation (indicator validity) to be too high, in a BST generated using this technique. Conversely, a classification that artificially inflates indicator heterogeneity within groups will yield a BST in which nuisance covariance is too high and indicator validity is too low.

To address the question of how well fallible classifications work in practice, along with a number of other questions, we performed a series of simulation studies. In Study 1, we examined the accuracy with which our data simulation algorithms reproduce important characteristics of a target data set. In Study 2, we

checked the utility of the bootstrap approach to interpreting taxometric curves. To perform a test of this approach unconfounded by fallible assignments of cases to groups, we provided the TaxSample program with the actual group membership of cases in taxonic data sets. In Study 3, we performed a more challenging test of the bootstrap approach. We obtained upper- and lower-bound estimates of the validity of structural inferences yielded by this approach, and compared the bootstrap approach with alternative tests that are popular in taxometric investigations.

Study 1: Precision and Bias in the Generation of Bootstrap Samples of Comparison Data

The aim of this study was to evaluate how well our algorithms reproduce important data parameters when generating a BSD or a BST.

Dimensional Data

For each of eight data conditions, 1,000 samples of target data with 4 indicators were generated. Indicators were correlated by virtue of shared loadings onto one or more latent dimensions (cf. Meehl & Yonce, 1994, 1996). For conditions in which indicator distributions were non-normal or non-continuous, an iterative technique was used to ensure that the desired level of indicator correlation was achieved (details are available upon request). For each condition, extensive checking revealed that the data generation program yielded target data sets with the intended indicator correlations and distributions. Default data parameters and population values were as follows: $N = 1{,}000$, one latent dimension, $r_{xy} = .50$, normal indicator distributions (skew = 0, kurtosis = 0). Condition codes indicate variations on these parameters. N600: $N = 600$; N300: $N = 300$; D2: two orthogonal dimensions ($r_{12} = r_{34} = .50$); S3: indicator distributions lognormal (a = 1, b = 1.67), skew = 3.00, kurtosis = 22.40; S6: indicator distributions lognormal (a = 1, b = 2.66), skew = 6.00, kurtosis = 105.76; C6: distributions initially normal, then cut into 6 ordered categories; D2S3C6: two orthogonal dimensions, distributions initially lognormal (a = 1, b = 1.67), then cut into 6 ordered categories. Within each condition, the SE of each statistic was estimated as the SD of values observed across the 1,000 target data sets (i.e., the 6,000 r and 4,000 M, SD, skew, and kurtosis values); for conditions with two latent dimensions, separate SEs were estimated for correlations between indicators loading onto the same dimension (r_{12} and r_{34}) vs. indicators loading onto different dimensions (r_{13}, r_{14}, r_{23}, and r_{24}).

For each target dimensional data set, one sample of bootstrap comparison data was generated using the DimSample program. Residuals were computed for each correlation and distributional moment. For each statistic x, the *Mdn* of residuals

($x_{comparison} - x_{target}$) was used to index bias, whereas the *Mdn* of absolute residuals was used to index precision (i.e., accuracy). The *Mdn* was used because many distributions of residuals and absolute residuals were skewed. Because the center and spread of these distributions varied widely across statistics and conditions, the ratio of a statistic's *Mdn* to its estimated SE (with signs dropped) was also calculated. Because a precision ratio of 1.00 means that the median error in the reproduction of a statistic equaled its standard error in the population, we consider precision ratios ≤ 1 to indicate acceptable results. We consider bias to be absent when it does not systematically differ from zero within or across data conditions. In cases of detectable bias; we consider it to be of a negligible magnitude to the extent that the bias ratio is small. Results for all conditions are presented in Table 1.

In every condition, indicator correlations and distributions were reproduced with good precision: All ratios of *Mdn* to SE were less than 1.00. There was no discernible bias in the reproduction of indicator correlations or Ms: Residuals were close to 0 and did not deviate systematically in either direction. However, biases were evident in the reproduction of higher distributional moments. The SD, skew, and kurtosis of reproduced distributions were smaller than in the target data sets. The ratios of *Mdn* to SE, however, suggest that the magnitude of each bias is very small. The common source of bias is likely to be a characteristic of the bootstrap resampling technique, not the other aspects of the algorithm within which we have embedded this technique. Specifically, when values are randomly resampled from an indicator distribution, the extent to which the most extreme values happen to be included in a bootstrap sample will influence its variance. Because there is no possibility of resampling values more extreme than those at the outer limits of an observed distribution, bootstrap samples that do not happen to include these most extreme observed values will tend to possess lower variance.

To examine the extent of the bias in bootstrap resampling, we generated univariate distributions of random unit normal values at each of several sample sizes. For each target sample, we calculated its SD, generated one bootstrap sample through random resampling with replacement, and calculated the SD of this bootstrap sample. Averaged across 10,000 replications at each sample size, the residual SDs were biased downward by an average of 7.3, 3.5, 1.8, 0.9, 0.4, 0.3, and 0.1% for $N = 10$, 20, 40, 80, 160, 320, and 640, respectively. This demonstration of downward bias in the variance of bootstrap samples may help to explain the other biases observed in Table 1. Given the positively skewed distributions used in several conditions, undersampling extreme values would be expected to reduce skew and kurtosis. Thus, the introduction of slight biases may be unavoidable when the bootstrap is used to reproduce observed distributions with sampling error. Fortunately, results presented in Table 1 suggest that the

TABLE 1
Study 1: Accuracy and Bias in the Reproduction of Indicator Correlations and Distributions in Dimensional Data Sets

Condition	*r*	*M*	*SD*	*Skew*	*Kurtosis*
			Precision		
N1000	.009 (.953)	.022 (.695)	.014 (.647)	.050 (.653)	.097 (.628)
N600	.011 (.933)	.027 (.665)	.019 (.652)	.065 (.651)	.120 (.618)
N300	.015 (.883)	.038 (.664)	.028 (.677)	.087 (.625)	.156 (.560)
D2	.004 (.143)	.022 (.682)	.015 (.659)	.051 (.658)	.098 (.649)
S3	.013 (.857)	.022 (.677)	.041 (.560)	.211 (.243)	1.855 (.116)
S6	.018 (.722)	.043 (.644)	.121 (.395)	.371 (.167)	4.803 (.082)
C6	.008 (.871)	.020 (.090)	.014 (.219)	.044 (.578)	.072 (.582)
D2S3C6	.008 (.311)	.013 (.098)	.025 (.169)	.186 (.120)	1.628 (.052)
			Bias		
N1000	.000 (.010)	.001 (.032)	−.001 (.044)	.000 (.002)	−.011 (.075)
N600	.000 (.029)	.000 (.003)	−.001 (.024)	.003 (.029)	−.020 (.101)
N300	.000 (.001)	.002 (.033)	−.002 (.061)	−.003 (.025)	−.033 (.118)
D2	.000 (.001)	.001 (.022)	−.001 (.059)	−.003 (.034)	−.010 (.068)
S3	.000 (.000)	.000 (.009)	−.003 (.037)	−.045 (.052)	−.410 (.026)
S6	−.001 (.025)	−.004 (.060)	−.015 (.050)	−.076 (.034)	−1.238 (.021)
C6	.000 (.016)	.001 (.005)	−.001 (.017)	−.003 (.035)	−.005 (.043)
D2S3C6	.000 (.015)	−.001 (.008)	−.003 (.020)	−.042 (.027)	−.358 (.011)

Default data parameters and population values were as follows: $N = 1{,}000$, one latent dimension, $r_{xy} = .50$, normal indicator distributions (skew = 0, kurtosis = 0). Condition codes indicate variations on these parameters. N600: $N = 600$; N300: $N = 300$; D2: two orthogonal dimensions ($r_{12} = r_{34} = .50$); S3: indicator distributions lognormal (a = 1, b = 1.67), skew = 3.00, kurtosis = 22.40; S6: indicator distributions lognormal (a = 1, b = 2.66), skew = 6.00, kurtosis = 105.76; C6: distributions initially normal, then cut into 6 ordered categories; D2S3C6: two orthogonal dimensions, distributions initially lognormal (a = 1, b = 1.67), then cut into 6 ordered categories. Within each condition, the *SE* of each statistic was estimated as the *SD* of values observed across the 1,000 target data sets. Bias is represented by the *Mdn* of residuals (comparison − target) for each statistic. Precision is represented by the *Mdn* of absolute residuals for each statistic. The *Mdn* was used because many distributions of residuals or absolute residuals were skewed. Because the center and spread of these distributions varied widely across statistics and conditions, values in parentheses show the *Mdn* divided by the estimated *SE* of a statistic for a particular data condition (with signs dropped).

magnitude of such biases is negligible: Considered as a fraction of the estimated *SE*, each *Mdn* bias was very small. This demonstration further suggests that the bias in variance decreases by about a factor of two with each doubling of N, reaching negligible levels by the time that sample sizes typical of taxometric studies are reached.

Taxonic Data

Because within-group indicator correlations and distributions are reproduced for taxonic data using the same technique already evaluated for dimensional data, we focused next on how well full-sample indicator correlations and indicator validities (group separation) are reproduced for taxonic data. Two sample sizes ($N = 300$ and 600) and four taxon base rates ($P = .50$, .25, .10, and .05) were crossed to yield 7 of 8 possible conditions; $N = 300$, $P = .05$ was omitted because the absolute size of the taxon would be unreasonably small. (An expanded range of data conditions was examined in Studies 2 and 3). For each condition, 1,000 samples of target data were generated with 4 indicators apiece; indicators were not systematically correlated within groups. Following the procedure of Meehl and Yonce (1994, 1996), a constant representing the desired level of indicator validity was added to each indicator for a predetermined proportion of cases representing the desired taxon base rate. Extensive checking revealed that the data generation program yielded target data sets with the intended characteristics. For each taxonic data set, one sample of bootstrap comparison data was generated using the TaxSample program. Residuals (comparison − target) and absolute residuals were computed for each correlation (in the full sample and within each group) and for indicator validity (group separation). Results are presented in Table 2.

Consistent with results shown in Table 1, indicator correlations within groups were reproduced with good precision and no discernible bias. The residuals for full-sample correlations and indicator validity were reproduced with good precision, but biased upward. Once again, these biases may be attributed to the tendency of bootstrap resampling to yield artificially low variance. In its standardized form (Cohen's d), indicator validity is calculated as the mean difference between groups relative to the pooled-variance estimate of the within-groups SD. Thus, reduced within-groups variance yields increased indicator validity. Likewise, indicators that distinguish groups with greater validity will be more highly correlated in the full sample. These biases appear to be of negligible magnitude: In all conditions, the ratio of Mdn to SE was very small.

Study 2: Using Bootstrap Samples of Comparison Data to Help Interpret Taxometric Curves

The results of Study 1 suggest that our simulation programs reproduce important data parameters with good precision and negligible bias. Study 2 was designed as a check on the proposed bootstrap approach itself, analogous to testing a new piece of experimental equipment under tightly controlled conditions before relying on its measurements in subsequent investigations. Using a much broader range of data conditions than Study 1, Study 2 tested the utility

TABLE 2
Study 1: Accuracy and Bias in the Reproduction of Indicator Correlations and Group Separations in Taxonic Data Sets

Sample Size (N)	*Taxon Base Rate (P)*	*Full-Sample r*	*Taxon r*	*Complement r*	*Indicator Validity (d)*
			Precision		
600	.50	.014 (.640)	.012 (.379)	.012 (.405)	.067 (.661)
600	.25	.015 (.641)	.018 (.426)	.011 (.442)	.074 (.659)
600	.10	.016 (.625)	.033 (.435)	.010 (.455)	.100 (.663)
600	.05	.016 (.634)	.058 (.520)	.010 (.435)	.125 (.653)
300	.50	.020 (.663)	.018 (.399)	.018 (.420)	.096 (.679)
300	.25	.021 (.652)	.027 (.424)	.014 (.397)	.107 (.678)
300	.10	.024 (.639)	.058 (.528)	.014 (.437)	.144 (.675)
			Bias		
600	.50	.000 (.015)	.000 (.003)	.000 (.004)	.003 (.028)
600	.25	.000 (.004)	.000 (.005)	.000 (.002)	.001 (.007)
600	.10	.000 (.010)	.001 (.017)	.000 (.002)	.005 (.033)
600	.05	.001 (.037)	.000 (.003)	.000 (.001)	.009 (.044)
300	.50	.002 (.072)	.000 (.003)	.000 (.002)	.015 (.103)
300	.25	.001 (.026)	.000 (.004)	.000 (.006)	.007 (.046)
300	.10	.001 (.016)	.000 (.001)	.000 (.004)	.009 (.041)

Bias is represented by the *Mdn* of residuals (comparison − target) for each statistic. Precision is represented by the *Mdn* of absolute residuals for each statistic. The *Mdn* was used because many distributions of residuals or absolute residuals were skewed. Because the center and spread of these distributions varied widely across statistics and conditions, values in parentheses show the *Mdn* divided by the estimated SE of a statistic for a particular data condition (with signs dropped).

of analyzing bootstrap samples of comparison data to help interpret taxometric curves.

Generating the Target Data Sets

To provide the cleanest test of interpretive utility, only taxonic target data sets were generated. The reason for this choice is that our approach calls for generating both BSDs and BSTs to help interpret a taxometric curve. Generating a BST requires the assignment of cases to groups, which can be done infallibly only by providing the algorithm with actual group membership. Because this information is only available where true groups exist, a pure test of the bootstrap approach requires taxonic target data sets. Study 3 will address this issue further, but for now we emphasize that Study 2 was designed to examine the *upper-bound* validity of structural inferences reached using empirical sampling distributions.

Only taxonic target data allowed us to do this in a manner unconfounded by the fallible assignment of cases to groups.

Default data parameters were as follows: $N = 1{,}000$, taxon base rate (P) = .50, indicator validity (d) = 2.00, within-group correlations (r) = .00, number of indicators (k) = 4, indicator skew (S) = 0, number of ordered categories (C) = 0 (i.e., continuous distributions). Each of these factors was varied, holding all other factors constant at the default level, using the following values: $N = 300, 400, 500, \ldots, 1{,}000$; $P = .05, .10, .15, \ldots, .50$; $d = 1.00, 1.25, 1.50, 1.75, 2.00$; $r = .00, .05, .10, \ldots, .30$; $k = 3, 4, 5, \ldots, 8$; $S = 0, 1, 2, \ldots, 6$ (for $S > 0$, lognormal distributions were used, with a = 1 and b selected to yield a population skew of S); $C = 6, 9, 12, 15, 0$ (values of 6–15 indicate the number of ordered categories; 0 indicates continuous distributions). This design includes a large number of factors that can influence the ability to detect taxonic structure, each of which varies widely across ranges of values that researchers might encounter. In each condition, 100 samples of taxonic target data were generated, and the data generation program was tested extensively prior to use in the study.

Bootstrapping Samples of Comparison Data

Bootstrap samples of comparison data were obtained for each target data set and used to generate empirical sampling distributions of taxometric curves. One practical question concerned the number of bootstrap samples to be generated and analyzed to flesh out an empirical sampling distribution. The bootstrapping literature suggests the convention of $B = 25$ to 200 bootstrap samples when estimating standard errors (Efron & Tibshirani, 1993; Mooney & Duval, 1993). In the present context, $B < 25$ may suffice because of the unusually large sample sizes ordinarily used in taxometric investigations (Haslam & Kim, 2002; Meehl, 1995). Because sampling error is likely to influence taxometric curves less than the results of many other analytic procedures, a smaller number of bootstrap samples may be required. For the present study, we chose $B = 10$ to balance the desirability of using as large a B as possible with the intensive computation required to generate and analyze the bootstrap samples of comparison data. In light of the iterative nature of our simulation algorithm (as well as the iterative factor analysis required within each iteration thereof), computing time is not a trivial consideration. We have found that $B = 10$ often works well, though further research is needed to chart the impact of varying B. As noted above, BSTs were generated by providing the TaxSample program with actual group membership.

Data Analysis

Our approach to generating empirical sampling distributions was designed for use with all taxometric procedures and consistency tests, and its logic applies

equally well to any of these. To test the approach in a way that allowed us to fully present results in the available space, we performed MAXEIG analyses of each data set using 50 windows that overlapped 90% with one another. It should be noted that holding the number of windows constant provides a conservative test of the identification of taxonic structure, as researchers ordinarily choose a number of windows judged to be most sensitive to detecting taxa in a particular analysis. For example, when a small taxon is suspected, investigators are advised to use a larger number of windows (Waller & Meehl, 1998). Rather than calculating the number of windows based on the sample size and taxon base rate, we chose to perform a more conservative test that cannot be accused of stacking the deck in favor of the successful detection of taxa.

Given the large number of curves that had to be compared in this study, we used an index to evaluate the utility of the bootstrap approach in interpreting taxometric curves. This index assesses the degree to which an averaged curve, obtained through analysis of research data, resembles curves in the sampling distributions of taxonic vs. dimensional comparison data. The similarity between two taxometric curves is quantified using the root mean square residual (RMSR) of the data points:

$$Fit_{\text{RMSR}} = \sqrt{\frac{\sum (y_{\text{res.data}} - y_{\text{boot.data}})^2}{N}}, \tag{5}$$

where $y_{\text{res.data}}$ refers to a data point on the curve for the research data, $y_{\text{boot.data}}$ refers to the corresponding data point on the averaged curve for either taxonic or dimensional bootstrap samples, and N refers to the number of points on each curve. Equation (5) is calculated twice, once to assess the fit of the research curves to the taxonic sampling distribution ($Fit_{\text{RMSR-tax}}$) and once to assess its fit to the dimensional sampling distribution ($Fit_{\text{RMSR-dim}}$). These two values are then integrated into a single comparison curve fit index (CCFI):

$$\text{CCFI} = \frac{Fit_{\text{RMSR-dim}}}{Fit_{\text{FMSR-dim}} + Fit_{\text{FMSR-tax}}} \tag{6}$$

CCFI values can range from 0 to 1, with lower values suggesting better fit for dimensional structure and higher values suggesting better fit for taxonic structure. The index is symmetric around .50 in that this middle value represents equivalent fit for both structures. It is important to note that the CCFI indexes the relative fit of taxonic and dimensional structural models, not the absolute goodness of fit of either model. In the present study, CCFI values above .50 represent correct detection of taxonic structure in the target data.

Results and Discussion

Table 3 summarizes the CCFI values. Because distributions were negatively skewed, the *Mdn* value is presented along with the first and third quartiles. When broken down by the levels of each factor, results are consistent with the expected influences of sample size, indicator validity, within-group correlations, and number of indicators: namely, the CCFI was larger with larger values of N, d, and k and with smaller values of r. Interestingly, taxon base rates were less influential, although the CCFI was largest at some of the smallest values of P, and increasing indicator skew yielded larger CCFIs. Notably, even when indicator distributions were cut into as few as 6 ordered categories, CCFIs remained at levels about as high as those obtained for continuous distributions.

In every analysis, the CCFI exceeded .50 and taxonic structure was identified correctly. This supports the effectiveness of our data simulation algorithms as well as the use of the CCFI as an interpretive aid. If there were problematic levels of bias or unacceptable precision in the reproduction of important data parameters, or if the CCFI functioned poorly, taxonic structure should not have been distinguishable from dimensional structure.

Study 3: A Comparative Analysis of the Interpretive Utility of Bootstrap Samples of Comparison Data

Study 2 extended the results of Study 1 and provided preliminary support for the validity of structural inferences based on bootstrap samples of comparison data. Study 3 was designed to address a number of critical questions that were not addressed by Study 2. First, how well does this approach work with dimensional data? We included both taxonic and dimensional target data sets in Study 3. Second, how well does this approach work when, as is the case in actual taxometric investigations, group membership is unknown? In Study 3, we obtained a lower-bound estimate of the validity of structural inferences (by using a minimally informed, fallible assignment of cases to groups) as well as an upper-bound estimate (by using actual group membership to generate BSTs). Third, how well does this approach work under more challenging data conditions than those in Study 2? Rather than varying one data parameter at a time while holding all others constant at favorable levels, we examined performance under more difficult conditions by simultaneously varying all data parameters. Fourth and finally, how well does this approach work compared to existing taxometric tests? An appropriately skeptical reader might wonder whether the generation and analysis of bootstrap samples of comparison data is worth the trouble. In Study 3, we evaluated the utility of the CCFI relative to several tests that have been used frequently in taxometric studies.

TABLE 3
Comparison Curve Fit Index (CCFI) Values

*Sample Size (**N**)*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1000*		
Q1	.792	.815	.838	.848	.860	.868	.876	.873		
Median	.823	.848	.863	.870	.877	.887	.895	.892		
Q3	.845	.864	.880	.892	.893	.899	.908	.907		
*Taxon Base Rate (**P**)*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*
Q1	.849	.890	.896	.894	.889	.884	.874	.879	.882	.873
Median	.867	.909	.912	.907	.906	.904	.891	.897	.899	.892
Q3	.888	.920	.925	.924	.919	.917	.910	.913	.912	.907
*Indicator Validity (**d**)*	*1.25*	*1.50*	*1.75*	*2.00*						
Q1	.706	.785	.840	.873						
Median	.734	.812	.860	.892						
Q3	.762	.833	.880	.907						
*Within-Group Corr. (**r**)*	*.00*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*			
Q1	.873	.866	.853	.836	.821	.809	.784			
Median	.892	.883	.874	.856	.844	.835	.818			
Q3	.907	.897	.890	.870	.866	.859	.842			
*Number of Indicators (**k**)*	*3*	*4*	*5*	*6*	*7*	*8*				
Q1	.817	.873	.903	.926	.932	.942				
Median	.843	.892	.919	.934	.941	.948				
Q3	.859	.907	.935	.943	.948	.956				
*Indicator Skew (**S**)*	*0*	*1*	*2*	*3*	*4*	*5*	*6*			
Q1	.873	.878	.889	.887	.904	.909	.913			
Median	.892	.897	.902	.906	.918	.925	.931			
Q3	.907	.915	.918	.921	.930	.935	.940			
*# Ordered Categories (**C**)*	*6*	*9*	*12*	*15*	*0*[a]					
Q1	.880	.877	.872	.874	.873					
Median	.894	.897	.891	.892	.892					
Q3	.910	.911	.906	.906	.907					

Within each condition, 100 samples of target data were generated. One factor was varied at a time, with default values for others as follows: $N = 1{,}000$, $P = .50$, $d = 2.00$, $r = .00$, $k = 4$, $S = 0$, $C = 0$ (continuous distributions). For each target data set, 10 bootstrap samples of comparison data were generated for each structure (taxonic and dimensional); actual group membership was used to generate taxonic comparison data. CCFI values above .50 are indicative of taxonic structure. In all conditions, 100% of the target data sets were correctly identified as taxonic.

[a]When $C = 0$, distributions were continuous.

Generating the Target Data Sets

A total of 10,000 taxonic and dimensional data sets (5,000 for each structure) were generated using a Monte Carlo design in which data parameters were independently randomly sampled from specified ranges. We drew values from ranges traditionally considered to be adequate for informative taxometric analyses (see Meehl, 1995). For taxonic data, random values were drawn for the following parameters of each target data set: sample size ($N = 300$ to 1,000), taxon base rate ($P = .10$ to .50), indicator validity ($d = 1.25$ to 2.00), within-group correlation ($r = .00$ to .30), indicator skew ($S = 0, 1, 2, \ldots, 6$), number of indicators ($k = 3, 4, 5, \ldots, 8$), and number of ordered categories ($C = 6$, 9, 12, 15, 0, with 0 representing continuous distributions). Values of N, P, d, and r were drawn from uniform, continuous distributions spanning the ranges listed above, whereas values of k and C were drawn from uniform distributions spanning the categories listed above. Values of S were drawn from the categories listed above, but to avoid the overrepresentation of high levels of skew—which may occur less frequently in research data than low to modest skew levels—S was determined at random with the following probabilities: 0 (.25), 1 (.20), 2 (.20), 3 (.15), 4 (.10), 5 (.05), and 6 (.05).

For dimensional data, values were drawn in the same way. However, because P, d, and r do not correspond to parameters of a dimensional model, they were combined to yield an expected indicator correlation using the following formula (Meehl & Yonce, 1994):

$$r_{xy} = \frac{P(1-P)d^2 + r}{P(1-P)d^2 + 1} \tag{7}$$

Extensive checking showed that our data generation programs created taxonic and dimensional target data sets with the intended indicator correlations and distributions.

Bootstrapping Samples of Comparison Data

For each target data set, 10 BSDs were generated. In addition, 10 BSTs were generated using each of several techniques for assigning cases to groups. First, for taxonic target data only, actual group membership was used to determine an upper-bound estimate of the validity of inferences of taxonic structure across a wider range of data conditions than was tested in Study 2. Second, for all target data sets, the base-rate classification technique was used to assign cases to groups. Specifically, the M estimate of the taxon base rate from the MAXEIG analysis was used to classify cases into the putative taxon (higher scoring) and complement (lower scoring) groups by rank-ordering cases according to their indicator total scores, then applying a threshold corresponding to the proportion

of the sample estimated to be taxon members. This yielded a lower-bound estimate of the validity of structural inferences. We consider this a lower-bound estimate because using only an estimated taxon base rate from the taxometric analysis does not call on any theoretical or empirical knowledge of the construct whose structure is being studied. The fact that this procedure can be fully automated reflects its failure to take advantage of any of the extra information that researchers often possess (e.g., a theoretically-derived base rate estimate, the proportion of cases scoring above a commonly-used threshold on a reliable and valid measure, a classification based on diagnostic criteria). Third, for all target data sets, cases were assigned to groups using Bayes' Theorem. This requires an estimate of the taxon base rate as well as the valid and false positive rates achieved by the optimal cutting score on each indicator, all of which can be obtained from taxometric analyses (Waller & Meehl, 1998).

Data Analysis

MAXEIG was performed on each data set, and the CCFI was calculated as in Study 2. To contextualize the performance of the bootstrap approach, all target data sets were also analyzed using three of the most popular quantitative indices in the taxometric literature (Haslam & Kim, 2002), none of which involves the analysis of comparison data.

Each MAXEIG curve can be used to calculate an estimate of the taxon base rate. Most descriptions of the taxometric method (e.g., Meehl, 1995; Waller & Meehl, 1998) assert that such estimates should cohere more closely around a single value for taxonic than dimensional latent structure. The SD of taxon base rate estimates across all curves in a full panel of results is a widely-used index of their consistency, so we calculated this for each target data set. Following the advice of Schmidt, Kotov, and Joiner (2002), a threshold of $< .10$ was used to indicate taxonic structure. In what follows, this test is abbreviated as SD.

Waller and Meehl (1998) introduced the goodness of fit index (GFI), which is well known in the structural equation modeling literature, to taxometrics. This index assesses the extent to which an observed indicator variance-covariance matrix can be reproduced by a predicted matrix. Specifically, a taxometric analysis is performed, the GCMT is used to estimate latent parameters, and these estimates are used to generate a predicted indicator variance-covariance matrix (Waller & Meehl, 1998). We calculated the GFI for each target data set and, following the lead of Waller and Meehl, used a threshold of $\geq .90$ to indicate taxonic structure.

Another popular consistency test is based on the shape of the distribution of Bayesian probabilities of taxon membership (Haslam & Kim, 2002). To generate this distribution, one begins by estimating the taxon base rate and the valid and

false positive rates achieved by the optimal cutting score on each indicator. Then, Bayes' Theorem is used to calculate each case's probability of belonging to the taxon, and the distribution of these probabilities is plotted. Probabilities are expected to cluster more closely around the extremes of 0 and 1 for taxonic than dimensional data (Waller & Meehl, 1998). To quantify this test, we used the technique introduced by Beauchaine and Beauchaine (2002): Calculate the proportion of Bayesian probabilities falling near 0 or 1 (specifically $< .10$ or $> .90$) and use a threshold of $\geq .80$ to indicate taxonic structure. In what follows, this test is abbreviated as PBayes.

In addition to evaluating each test using the thresholds listed above, we examined its discriminating power using a measure independent of threshold: the area under a receiver operating characteristic (ROC) curve. This area (A) represents the probability that a randomly chosen pair of taxonic and dimensional target data sets would be correctly distinguished (Hanley & McNeil, 1982). For example, if $A = .80$ for the GFI, this means that a randomly chosen taxonic data set is 80% likely to yield a higher GFI value than a randomly chosen dimensional data set.

Results and Discussion

The classification accuracy for each of the four tests in analyses of all 10,000 target data sets is shown in Table 4. Using the recommended thresholds, the SD, GFI, and PBayes tests correctly classified the latent structure of 51.0, 68.0, and 49.5% of all data sets, respectively. Because these poor discriminations may stem from poorly chosen thresholds, ROC analyses were performed to examine performance independent of threshold as well as to evaluate performance at a threshold optimized for this sample of data sets; these results also appear in Table 4. The SD, GFI, and PBayes tests yielded $A = .381$, .741, and .608, respectively. Although these analyses suggested that the recommended thresholds for the SD and GFI tests may be too low, using thresholds optimized for this sample yielded little improvement in the accuracy of these tests. The recommended threshold for PBayes was much too high for this sample, and lowering it to .510 improved the test's classification accuracy from 49.5% to 65.6%.

Overall, results were considerably stronger for the CCFI. Using BSTs generated with a Bayesian classification of cases, the CCFI achieved 76.7% accuracy; its sensitivity to taxonic structure was good (89.9%), but its specificity was poor (63.5%). Because the Bayesian classification of cases yielded such weak results for dimensional data, and because alternative classification approaches are available, we do not discuss this technique further. Overall classification accuracy was better (84.3%) when BSTs were generated using the base-rate classification technique, with somewhat lower sensitivity (77.8%) than specificity (90.8%).

TABLE 4
Study 3: Discriminating Power of Each Taxometric Test

	SD of Base Rate Estimates	*GFI*	*Proportion of Extreme Bayesian Probabilities*	*CCFI*[a]	*CCFI*[b]	*CCFI*[c]
Recommended threshold	.10[d]	.90[e]	.80[e]	.50[e]	.50[e]	.50[e]
Percent correct:						
All data	51.0	68.0	49.5	—	84.3	76.7
Taxonic data	79.3	82.9	38.0	99.3	77.8	89.9
Dimensional data	22.7	53.2	60.9	—	90.8	63.5
ROC analyses:						
Area under curve	.381	.741	.608	—	.932	.888
Optimal threshold	.150[d]	.928[e]	.510[e]	—	.355[e]	.672[e]
Percent correct at optimal thresh.	53.7	69.6	65.6	—	86.2	78.5
Logistic regressions:						
Percent correct	55.7	68.8	61.9	—	85.7	78.2
Incremental validity[f]	1.9	.1	−0.1	—	—	—

GFI = goodness of fit index. CCFI = comparison curve fit index.

[a]Calculated for bootstrap samples of taxonic data generated using actual group membership.

[b]Calculated for bootstrap samples of taxonic data generated using base-rate classification of cases to groups.

[c]Calculated for bootstrap samples of taxonic data generated using Bayes' Theorem classification of cases to groups.

[d]Values below this threshold are indicative of taxonic structure.

[e]Values above this threshold are indicative of taxonic structure.

[f]Calculated as the percent correct for the logistic regression using this test plus the CCFI minus the percent correct for the logistic regression using only the CCFI.

ROC analysis yielded $A = .932$ for the CCFI, a value considerably higher than that for the SD, GFI, and PBayes tests.

As argued earlier, we believe that results based on the base-rate classification technique represent a lower-bound estimate of the validity of structural inferences using the bootstrap. When generating BSTs, researchers should be able to assign cases to groups in at least as valid a manner as this. At the other extreme, an upper-bound estimate was obtained by using actual group membership to generate BSTs for taxonic target data sets. Taxonic structure was correctly identified by the CCFI in 99.3% of the 5,000 taxonic target data sets. Thus, in tests unconfounded by the fallibility of case classification, the bootstrap approach identified taxonic structure with near-perfect accuracy. Ultimately, the validity that one will achieve in practice depends on how well cases can be assigned to groups. Under conditions similar to those studied here, researchers could expect to achieve at least the lower-bound accuracy of 77.8% in the identification of

taxonic structure. To the extent that they can assign cases to groups more effectively, this figure could be pushed toward the upper-bound accuracy of 99.3%.

As a final examination of the extent to which each test provides useful information, we performed logistic regression analyses to examine the validity of inferences based on multiple tests as well as the incremental validity of each of the three conventional tests when used as an adjunct to the CCFI. The second-to-last row of Table 4 shows the classification accuracy of separate logistic regressions in which each test served as the sole predictor. Like the ROC analyses, these results reveal that when the logistic regression model was allowed to determine its own threshold, accuracy improved minimally for the SD and GFI tests and more substantially for the PBayes test relative to the use of recommended thresholds. More important are the logistic regressions performed using multiple tests as predictors. First, we included the SD, GFI, and PBayes tests in a single analysis. This yielded 72.4% correct classifications, still far short of the rate achieved using the CCFI alone (85.7%). Second, we examined the incremental validity of the SD, GFI, and PBayes tests by simultaneously entering one test at a time along with the CCFI into three logistic regression analyses. As shown in the bottom row of Table 4, the SD test added 1.9% to classification accuracy, whereas accuracy was relatively unaffected by adding either the GFI or PBayes test (+0.1% and −0.1%, respectively). Thus, all three conventional tests combined fail to attain accuracy as high as the CCFI, and none of the three tests contributes much incremental validity over and above the CCFI.

These global analyses provide compelling support for the bootstrap approach, but do not explore the possibility that alternative tests may be superior under certain data conditions. To evaluate the robustness of the CCFI's comparatively strong performance, more fine-grained analyses were performed using subsets of the 10,000 target data sets that varied along each data parameter. In all analyses, the CCFI was generated using the base-rate classification technique to provide a conservative test of the bootstrap approach. Figure 1 shows the performance of the CCFI and the three conventional tests when the sample is broken down by each factor; discriminating power is indexed using A, with error bars extending to ± 1 *SE* of A (calculated using the formula in Hanley & McNeil, 1982). The most striking finding is that under all conditions, the CCFI performed substantially better than the SD, GFI, and PBayes tests. The discrepancies were often quite large, with the discriminating power of the alternative tests seldom approaching that of the CCFI. In fact, with one exception—indicator skew—the poorest performance of the CCFI (its lowest A across all levels of a given factor) exceeded the best performance of every other test (its highest A for that factor). Thus, the superiority of the CCFI to the other tests studied here appears highly robust across data conditions. We briefly review the performance of each test across levels of each factor below.

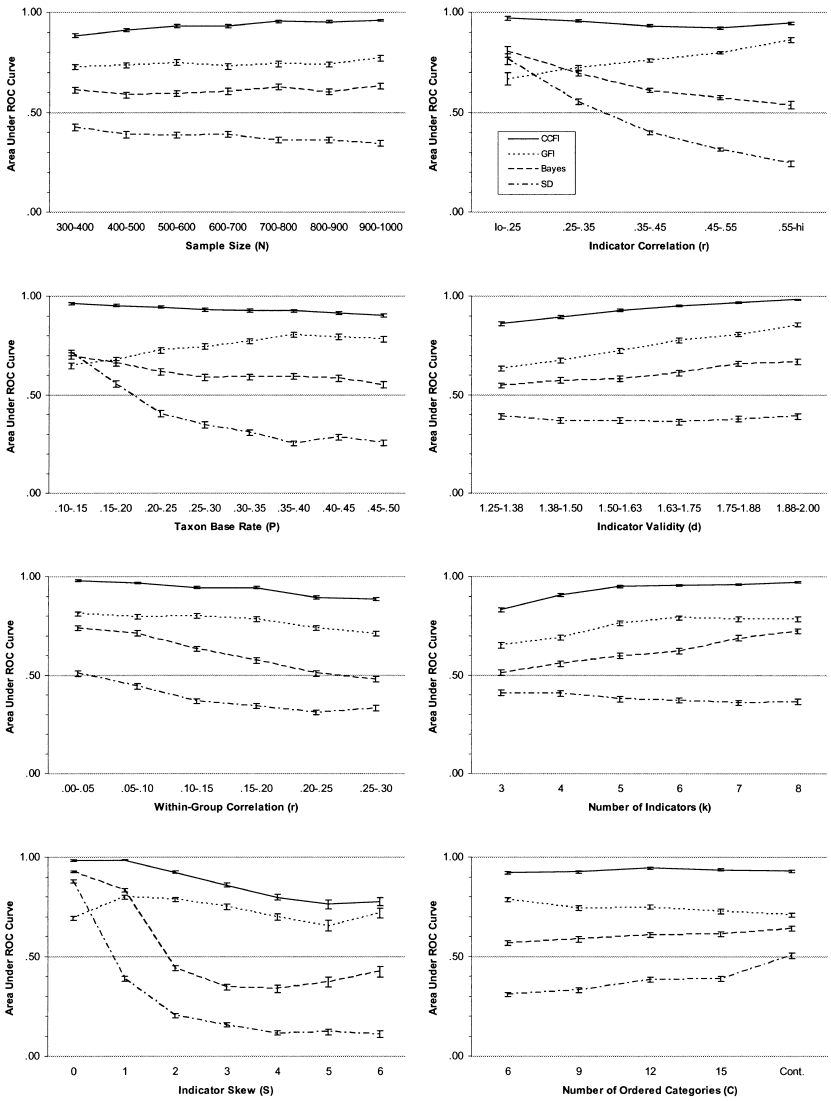

FIGURE 1 Validity with which each of four tests distinguished taxonic from dimensional structure in Study 3. Discriminating power is indexed independent of threshold by using the area under the ROC curve (A), with error bars representing $\pm$ 1 SE of A (calculated using the formula in Hanley & McNeil, 1982). Each graph contains the results across levels of one data parameter that varied in this study. CCFI = Comparison Curve Fit Index; GFI = Goodness of Fit Index; Bayes = proportion of cases for which the Bayesian probability of taxon membership was less than .10 or greater than .90; SD = standard deviation of taxon base rate estimates.

As either sample size or the number of indicators increased, so did the performance of all but the SD test. Because the SD test yielded results worse than chance (i.e., $A < .50$), providing larger samples or more indicators merely amplified its poor performance. Our finding that the SD test would have performed better if SDs above, rather than below, the threshold were used to infer taxonic structure—when $A < .50$, reversing the inferences for values above vs. below the threshold would yield $A > .50$—is consistent with more extensive results on base rate consistency (J. Ruscio, in press).

All four tests were sensitive to changes in (full-sample) indicator correlations (which are a function of the values of P, d, and r for taxonic data; see Equation (7)). Whereas the discriminating power of the CCFI dropped slightly as correlations increased, the power of the SD and PBayes tests dropped much more substantially. In contrast, the GFI discriminated latent structures better with larger indicator correlations.

As in Study 2, the CCFI performed slightly better with lower taxon base rates. All three of the other tests performed similarly to each other with low taxon base rates. As base rates increased, the performance of the GFI improved, the performance of PBayes declined a bit, and the performance of the SD test declined substantially.

Increasing indicator validity improved the performance of all tests (with the possible exception of the SD test), as did decreasing within-group indicator correlations.

Indicator skew had some of the most dramatic effects on discriminating power. Whereas the CCFI, SD, and PBayes tests performed very well when $S = 0$, among these tests only the CCFI was able to maintain relatively good performance with increasing skew. Curiously, the GFI was the weakest test when $S = 0$ and its performance followed a strange trajectory when $S > 0$: its discriminating power increased with mild skew ($S = 1$), tapered off as S increased to 5, then rose again at $S = 6$. This erratic pattern is difficult to interpret, but the small SEs associated with this large sample suggest that something more than sampling error is responsible for the finding.

GENERAL DISCUSSION

Results of these three studies suggest that our data simulation algorithms reproduce important characteristics of target data sets with good precision and negligible bias, and that analyses of these simulated comparison data can aid in the interpretation of taxometric curves. Study 1 showed that indicator correlations and distributional moments in bootstrap samples of comparison data reproduce those in target data sets at least as closely as if a new sample had been drawn from the same population. The use of bootstrap resampling to reproduce

indicator distributions does introduce some biases, but the magnitude of these biases appears to be very small, especially with the large sample sizes typical of taxometric investigations. Studies 2 and 3 did not directly examine precision or bias, but the good performance of the CCFI hints at the adequacy with which our simulation algorithms reproduce important data parameters. When the cleanest test of the algorithms and the CCFI was performed—by generating BSTs using actual group membership, rather than introducing the confound of fallible case classification; taxonic structure was identified with 100% accuracy in Study 2 and with 99.3% accuracy in the more challenging conditions of Study 3. These accuracy levels suggest that the biases identified in Study 1 are indeed negligible and that precision is quite satisfactory for the purpose at hand.

Results further revealed that the CCFI distinguishes taxonic and dimensional structure with greater accuracy than three tests that are frequently used in taxometric investigations (Haslam & Kim, 2002). In Study 3, ROC analyses revealed that the SD, GFI, and PBayes tests would correctly distinguish a randomly chosen pair of taxonic and dimensional data sets with 38.1, 74.1, and 60.8% accuracy, respectively. By comparison, a randomly chosen taxonic data set would yield a larger CCFI value than a randomly chosen dimensional data set 93.2% of the time. Because this figure was obtained by generating BSTs using the minimally informed base-rate classification technique, we believe it represents a lower-bound estimate of what researchers can expect to achieve under similar data conditions. Even using this coarse case classification strategy, the CCFI performed significantly better than all three conventional tests at each level of each factor examined in Study 3. In sum, the CCFI appears to discriminate between taxonic and dimensional latent structures more powerfully than three popular tests under a wide range of data conditions considered acceptable for taxometric studies.

The present findings should be evaluated with several important issues in mind. First, we have argued that applying bootstrap methods to taxometrics may provide a more appropriate framework for interpretation than relying exclusively on the results of Monte Carlo studies. While Monte Carlo studies typically include a small slice of the full parameter space and little, if any, variation in the implementation of an analytic procedure, the bootstrap tailors simulations to parameters estimated from the data and implements analyses as desired in a particular study. For these reasons, we believe there to be great potential value in applying the bootstrap to the taxometric method. The present studies do not conclusively demonstrate that this potential can be realized for procedures other than MAXEIG, as due to space constraints we focused on using a single procedure to study and illustrate the bootstrap approach. Nonetheless, we see no reason why the utility of the bootstrap approach should be unique to MAXEIG analyses, and believe that the logic of the approach will generalize to the other procedures in the taxometric method. We encourage researchers to examine the

utility of the bootstrap as an interpretive aid for the full range of taxometric procedures and consistency tests.

Second, the present results suggest that the bootstrap might allow researchers to make informed judgments about the adequacy of their data for taxometric analyses. In the absence of comparative taxonic and dimensional results, one must evaluate the adequacy of a data set one estimated model parameter at a time, often in the face of ambiguous or contradictory information. For example, some parameters may be judged acceptable for taxometric analysis, whereas others may not. Or, all parameters may seem adequate, but just barely; in which case reaching a judgment about such a unique configuration can be difficult because the joint influence of marginally acceptable parameters is poorly understood. Moreover, implementing any taxometric procedure requires that one make a number of potentially important choices from an array of analytic options, yet virtually no guidelines exist for the most appropriate way to do so. Submitting bootstrap samples of comparison data to a taxometric procedure, implemented in a particular way, provides the investigator with valuable and otherwise unavailable information about the expected results under the taxonic and dimensional models. This could be useful in judging the adequacy of a unique sample of research data for a taxometric analyses implemented in particular ways.

Third, realizing the potential of this approach requires the generation of bootstrap samples of taxonic and dimensional comparison data in which important data parameters are reproduced well. Although our simulation algorithms appear promising in this respect, this does not preclude further improvements. For example, our algorithms are not model-based; they do not attempt to specify the process giving rise to the observed data, only to reproduce indicator correlations and distributions in ways that involve either two latent groups or one or more latent dimensions. We present our simulation algorithms in the spirit of programs that address a pragmatic need with satisfactory results, but we make no claim that they represent an optimal solution to the problem of generating comparison data for taxometric analyses.

Fourth, using our algorithm to generate BSTs requires the assignment of cases to putative groups. A requirement of this sort is not unique to our approach, as some understanding of group membership is necessary for any effort to estimate and reproduce data parameters within groups. In many research contexts, investigators may have only an educated guess about the relative sizes of groups or the class membership of each case. When more theoretically-based classification schemes are not available, one can construct a criterion using an estimate of the base rate alone, as was done in Study 3. However, we view this as a method of last resort, and believe that researchers can do better by taking advantage of additional information. The gap between the upper- and lower-bound estimates of accurate taxon detection in Study 3 shows that the utility of the bootstrap approach is enhanced by better case classification, underscoring the need for

further research on how best to assign cases to groups. In the meantime, it is noteworthy that even the lower-bound estimate of the validity of structural inferences achieved by the bootstrap exceeded that achieved by three other tests commonly used in taxometric studies. On a related note, it is critical to bear in mind that the failure of an investigator to provide an appropriate case classification does not impugn the effectiveness of a data simulation algorithm. The TaxSample program reproduces indicator correlations, distributions, and validities with good precision and negligible bias—*relative to the case classification that is provided*. In other words, the performance of the simulation algorithm itself must be evaluated against the target data set that is supplied by the user. The standard warning of "garbage in, garbage out" applies here; one cannot reasonably hold a simulation algorithm accountable for a poor criterion provided by an investigator.

Fifth, the CCFI represents one way that empirical sampling distributions of results can be compared with obtained taxometric results to facilitate interpretation. We relied on this index in our simulation studies because the sheer number of target and comparison curves that were generated precluded visual curve inspection by knowledgeable raters. In addition, we wanted to use an objective index whose mechanical nature would remove the possibility of experimenter bias in the interpretation of taxometric curves. The results of Study 3 suggested that the accuracy of the CCFI is at least as impressive as that achieved by three popular alternative tests and may well surpass them under conditions typical of taxometric investigations. Despite this initial success, we believe that caution should be exercised if the CCFI is used in taxometric studies. One reason is that we have not established that the CCFI yields more accurate structural inferences than visual comparison of research and bootstrap curves by knowledgeable users of the taxometric method. Human beings can be remarkably adept at pattern recognition, and there is an inevitable loss of information inherent in calculating a quantitative index to summarize something as complex and nuanced as the similarity or dissimilarity of taxometric curves. Additional research is required to determine whether interpretations based on visual inspections by informed researchers are more or less valid than those based on the CCFI. Another reason for caution is that the CCFI currently requires the averaging of curves. This may be appropriate when all indicators are sufficiently valid, but when one or more indicators is not, averaged results may obscure important information and obfuscate interpretation. Although the CCFI could be calculated separately for each curve in a full panel, research would need to determine whether this improves the validity of structural inferences and, if so, under what conditions.

Two additional caveats are important for understanding what the bootstrap can and cannot accomplish for taxometric analyses. As noted earlier, there is no guarantee that any set of research data will yield results that unambiguously support an inference of taxonic or dimensional structure. Empirical sampling

distributions representing taxonic and dimensional structure may overlap considerably, or the results for the research data may be ambiguous in appearing equally consistent, or inconsistent, with both distributions. Ambiguous results may stem from inadequate data, a poorly chosen or implemented analysis, or poor correspondence between actual latent structure and the taxonic and dimensional structural models. Although the bootstrap does not guarantee easily interpretable results, it offers the benefit of alerting researchers to ambiguous and potentially misleading results that might otherwise lead to unwarranted structural inferences. In this way, consulting empirical sampling distributions for taxonic and dimensional structures may prevent overconfident conclusions.

Another important caveat is that applying the bootstrap to taxometrics does not reduce sampling error or alleviate methodological artifacts in the original sampling of cases for investigation. For example, even if the structure of a latent variable actually is taxonic in a specified population, a particular sample of data might produce dimensional results due to sampling error. Bootstrap samples of comparison data will not counteract such bad luck, as this source of sampling variation is not taken into account in the generation of empirical sampling distributions. Likewise, systematic biases stemming from an inappropriate sampling scheme might yield misleading results, and the analysis of comparison data will not identify such problems.

Although our iterative algorithm was devised for use in taxometric research, it may be possible to adapt it for use with other analytic procedures. By systematically varying the number of latent classes, one could generate bootstrap samples to help interpret the results of cluster analyses, mixture models, latent class analyses, or other analyses when these procedures are used to determine the number of latent classes. Likewise, by systematically varying the number of latent factors, rather than allowing the simulation algorithm to determine this number for itself, one could generate bootstrap samples to help interpret the results of exploratory or confirmatory factor analyses. Future research is needed to ascertain whether bootstrap methodology may be used to improve the validity of structural inferences reached using data analytic procedures outside the taxometric method.

We would also like to draw attention to an important and versatile feature of bootstrap methodology known as *adaptive estimation* (Efron & Tibshirani, 1993). In principle, this involves letting the data suggest how data analyses should be implemented. By performing a wide range of potentially informative tests in varying ways, and examining the extent to which empirical sampling distributions diverge for each test, one can identify empirically the analytic approach that most powerfully distinguishes taxonic and dimensional structure. For example, adaptive estimation could be used to determine whether indicators constructed on the basis of different criteria yield differentially informative results, or whether nonlinear transformations applied to reduce indicator skew

influence the power of an analysis to distinguish the structural models. Similarly, given indications in Study 3 that the interpretive thresholds recommended for several taxometric consistency tests may not be located optimally (or even that generally optimal thresholds may not exist), adaptive estimation might be used to determine the analyses that best differentiate taxonic and dimensional structure and then to set thresholds where the empirical sampling distributions intersect. Particularly if Monte Carlo studies do not establish generally applicable thresholds, threshold values might be allowed to vary from analysis to analysis in a way that takes into consideration both the unique characteristics of the research data and the procedural implementation selected. Future research is needed to determine whether the technique of adaptive estimation can be used to advantage in taxometric analyses.

In the not-so-distant past, performing the complex data analyses involved in classification research was demanding. For the average researcher, it would have been impractical to complicate matters further by generating empirical sampling distributions for interpretive purposes. However, given advances in computing power and the sophistication of software for data analysis, such distributions can now be obtained to test the performance of a taxometric procedure or consistency test for a particular set of research data. Analyses of bootstrap samples of comparison data supplement Monte Carlo studies that cannot realistically include all data conditions and procedural implementations that researchers might encounter. By examining sample-specific comparative results, investigators may reach more accurate structural inferences, hold better-justified levels of confidence in those inferences, and more readily identify instances when no structural inferences should be drawn. The simulation studies presented here suggest that investigators have little to lose and much to gain by applying the bootstrap to taxometrics.

ACKNOWLEDGMENTS

This work was supported in part by National Research Service Award MH12675 to A. M. Ruscio from the National Institute of Mental Health.

REFERENCES

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Beauchaine, T. P., & Beauchaine, R. J. (2002). A comparison of maximum covariance and k-means cluster analysis in classifying cases into known taxon groups. *Psychological Methods, 7,* 245–261.

Cronbach, L. J., & Meehl, P. E. (1995). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. San Francisco: Chapman & Hall.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

Grove, W. M. (2004). The MAXSLOPE taxometric procedure: Mathematical derivation, parameter estimation, consistency tests. *Psychological Reports, 95,* 517–550.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Haslam, N., & Kim, H. (2002). Categories and continua: A review of taxometric research. *Genetic, Social, and General Psychology Monographs, 128,* 271–320.

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods, 3,* 91–103.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60,* 117–174.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–274.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology, 113,* 39–43.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74,* 1059–1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports, 78,* 1091–1227.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Murphy, E. A. (1964). One cause? Many causes? The argument from the bimodal distribution. *Journal of Chronic Disease, 17,* 301–324.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Erlbaum.

Ruscio, J. (in press). Taxometric analysis: An empirically–grounded approach to implementing the method. *Criminal Justice and Behavior*.

Ruscio, J., & Ruscio, A. M., & Keane, T. M. (2004). Using taxometric analysis to distinguish a small latent taxon from a latent dimension with positively skewed indicators: The case of Involuntary Defeat Syndrome. *Journal of Abnormal Psychology, 113,* 145–154.

Schmidt, N. B., Kotov, R., & Joiner, T. E., Jr. (2004). *Taxometrics: Toward a new diagnostic scheme for psychopathology*. Washington, DC: American Psychological Association.

Stone, C. A. (2000). Monte Carlo based null distributions for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58–75.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage.

Waller, N. G., Underhill, J. M., & Kaiser, H. A. (1999). A method for generating simulating plasmodes and artificial test clusters with user-defined shape, size, and orientation. *Multivariate Behavioral Research, 34,* 123–142.

Wichura, M. J. (1988). Algorithm AS 241: The percentage points of the normal distribution. *Applied Statistics, 37,* 477–484.

APPENDIX

Simulation Program Code Written in R

The functions "DimSample" and "TaxSample" simulate dimensional and taxonic comparison data, respectively. The former calls the "Factor.Analysis" function, which is available upon request. Note that for taxonic simulations, the final column in the supplied data must contain a criterion variable that classifies each case (1 = complement, 2 = taxon). R can be downloaded for free at http://cran.r-project.org/mirrors.html, and our program code (accompanied by a suite of taxometric programs that integrate the simulation and analysis of comparison data into the rubric of taxometric analyses, with an extensive documentation file) can be downloaded from http://www.taxometricmethod.com.

```
##############################################################################
DimSample <- function(x, Group = 0, Trials = 5, Multiplier = 1, seed = 1)
{
# Read N (# cases) and k (# indicators), bootstrap distribution for each
  indicator
       x <- as.matrix(x)
       N <- dim(x)[1]
       k <- dim(x)[2]
       Freq.Dist <- matrix(nrow = N, ncol = k)
       for (i in 1:k)
               Freq.Dist[,i] <- sort(sample(x[,i], replace = T))

# Compute target correlation matrix and store copy for desired correlations
       Target.Corr <- cor(x[,1:k])
       Desired.Corr <- Target.Corr

# Determine number of latent factors to use and generate random normal data
  for shared and unique components
       Factors <- sum(eigen(Desired.Corr)$values > 1)
       Shared.Comp <- matrix(rnorm(N * Factors, mean = 0, sd = 1), nrow =
                                N, ncol = Factors)
       Unique.Comp <- matrix(rnorm(N * k, mean = 0, sd = 1), nrow = N,
                                ncol = k)
       Unique.Load <- vector("numeric", k)
       Loadings <- matrix(nrow = k, ncol = Factors)

# Create empty matrix for simulated data and initialize variables prior to
  iterations
       y <- matrix(0, nrow = N, ncol = k)
       Iter <- 0
       Best.RMSR <- 1
       j <- 0
```

```
# Begin loop that ends when specified number of iterations pass without
  improvement in correlational reproduction
       while (j < Trials)
       {
               Iter <- Iter + 1

# Calculate factor loadings and apply to reproduce desired correlations
                Fact.Anal <- Factor.Analysis(Desired.Corr, Corr.Matrix = T,
                                             N.Factors = Factors)
                if (Factors == 1) Loadings[,1] <- Fact.Anal$loadings
                       else for (i in 1:Factors)
                                    Loadings[,i] <- Fact.Anal$loadings[,i]
                Loadings[Loadings > 1] <- 1
                Loadings[Loadings < -1] <- -1
                if (Loadings[1,1] < 0) Loadings <- Loadings * -1
                for (i in 1:k)
                        if (sum(Loadings[i,] ^ 2) < 1)
                                Unique.Load[i] <- sqrt(1 -
                                                  sum(Loadings[i,] ^ 2))
                        else Unique.Load[i] <- 0
                for (i in 1:N)
                       for (l in 1:k)
                              y[i,l] <- sum(Shared.Comp[i,] *
                        Loadings[l,]) + Unique.Comp[i,l] * Unique.Load[l]

# Replace normal distributions with bootstrapped distributions
               for (i in 1:k)
               {
                       y <- y[sort.list(y[,i]),]
                       y[,i] <- Freq.Dist[,i]
               }


# Calculate residual correlations (target - reproduced) and check to see if
  best match yet
                Reproduced.Corr <- cor(y)
                Residual.Corr <- Target.Corr - Reproduced.Corr
                RMSR <- sqrt(sum(Residual.Corr[lower.tri(Residual.Corr)]^2) /
                                 length(Residual.Corr[lower.tri(Residual.
                                        Corr)]))
                if (RMSR < Best.RMSR)
                {
                       Best.RMSR <- RMSR
                       Best.Corr <- Desired.Corr
                       Best.Res <- Residual.Corr
                       Desired.Corr <- Desired.Corr + Multiplier *
                            Residual.Corr
                       j <- 0
                }
                else
```

```
                {
                        j <- j + 1
                        Mult <- Multiplier / (2 ^ j)
                        Desired.Corr <- Best.Corr + Mult * Best.Res
                }
        }

# Report the RMSR correlation for the best correlational reproduction
        Iter <- Iter - Trials
        if (Group == 0) cat("\nDimensional data set:  N =",N,", RMSR r =",
                                                round(Best.RMSR,3),"\n")
        if (Group == 1) cat("\n         Complement:  N =",N,", RMSR r =",
                                                  round(Best.RMSR,3))
        if (Group == 2) cat("\n              Taxon:  N =",N,", RMSR r =",
                                                  round(Best.RMSR,3))

# Return the data set that best reproduced the correlations
        Fact.Anal <- Factor.Analysis(Best.Corr, Corr.Matrix = T, N.Factors
                                    = Factors)
        if (Factors == 1) Loadings[,1] <- Fact.Anal$loadings
                else for (i in 1:Factors)
                         Loadings[,i] <- Fact.Anal$loadings[,i]
        Loadings[Loadings > 1] <- 1
        Loadings[Loadings < -1] <- -1
        if (Loadings[1,1] < 0) Loadings <- Loadings * -1
        for (i in 1:k)
                 if (sum(Loadings[i,] ^ 2) < 1)
                         Unique.Load[i] <- sqrt(1 - sum(Loadings[i,] ^ 2))
                 else Unique.Load[i] <- 0
        for (i in 1:N)
                 for (l in 1:k)
                         y[i,l] <- sum(Shared.Comp[i,] * Loadings[l,]) +
                                           Unique.Comp[i,l] * Unique.Load[l]
        y <- apply(y, 2, scale)
        for (i in 1:k)
        {
                y <- y[sort.list(y[,i]),]
                y[,i] <- Freq.Dist[,i]
        }
        return(y)
}

###############################################################################
TaxSample <- function(x, Trials = 5, Multiplier = 1)
{
# Read N (# cases) and k (# indicators)
        x <- as.matrix(x)
        N <- dim(x)[1]
        k <- dim(x)[2] - 1
```

```
# Select subsamples of cases belonging to complement (x1) and taxon (x2)
       x1 <- x[(x[,k + 1] == 1),]
       x2 <- x[(x[,k + 1] == 2),]

# Generate simulated data for complement and taxon by calling DimSample
  twice
       tax <- DimSample(x2[,1:k], Group = 2, Trials, Multiplier)
       com <- DimSample(x1[,1:k], Group = 1, Trials, Multiplier)

# Create matrix for simulated taxonic data set and merge complement and
  taxon
       y <- matrix(0, nrow = N, ncol = k + 1)
       y[,1:k] <- rbind(tax,com)

# Calculate the RMSR correlation in the full sample
       Residual.Corr <- cor(x[,1:k]) - cor(y[,1:k])
       RMSR <- sqrt(sum(Residual.Corr[lower.tri(Residual.Corr)]^2) /
                  length(Residual.Corr[lower.tri(Residual.Corr)]))
       cat("\n    Taxonic data set:  N =",N,", RMSR r =",round(RMSR,3),
                                     "\n")

# Add a column to identify complement (1) and taxon (2) members and return
  the data set
       y[,k + 1] <- c(rep(2, dim(x2)[1]), rep(1, dim(x1)[1]))
       return(y)
}
```