Psychological Assessment
2009, Vol. 21, No. 4, 578–594

1040-3590/09/$12.00 DOI: 10.1037/a0016558

Using Comparison Data to Differentiate Categorical and Dimensional Data by Examining Factor Score Distributions: Resolving the Mode Problem

John Ruscio
The College of New Jersey

Glenn D. Walters
Federal Correctional Institution, Schuylkill

Factor-analytic research is common in the study of constructs and measures in psychological assessment. Latent factors can represent traits as continuous underlying dimensions or as discrete categories. When examining the distributions of estimated scores on latent factors, one would expect unimodal distributions for dimensional data and bimodal or multimodal distributions for categorical data. Unfortunately, identifying modes is subjective, and the operationalization of counting local maxima has not performed very well. Rather than locating and counting modes, the authors propose performing parallel analyses of categorical and dimensional comparison data and calculating an index of the relative fit of these competing structural models. In an extensive Monte Carlo study, the authors replicated prior results for mode counting and found that trimming distributions' tails helped. However, parallel analyses of comparison data achieved much greater accuracy, improved base rate estimation, and afforded consistency checks with other taxometric procedures. Two additional studies apply this approach to empirical data either known to be categorical or presumed to be dimensional. Each study supports this new method for factor-analytic research on the latent structure of constructs and measures in psychological assessment.

Keywords: factor analysis, categories, dimensions, taxometrics, parallel analysis

The field of psychological assessment has a long history of factor-analytic research for the purpose of examining the latent structure of its constructs. Typically, this has involved using exploratory or confirmatory factor analyses to determine the number of distinct factors assessed by a measure (Brown, 2006; Gorsuch, 1983). Even though most users of factor analysis believe that a latent factor represents a trait as a continuous underlying dimension, this need not be the case. In two editions of an important early text on factor analysis, Thurstone (1935, 1947) argued that factors could represent traits categorically, taking on two or more discrete values only. This possibility has received increased attention as investigators have recognized the importance of distinguishing between categorical and dimensional latent structures for the development, evaluation, and use of assessment instruments (Meehl, 1992; J. Ruscio, Haslam, & Ruscio, 2006). In addition to improving research design and statistical power (MacCallum, Zhang, Preacher, & Rucker, 2002), assessing categorical constructs versus dimensional constructs promotes divergent goals and strategies. For a categorical construct, one assigns individuals to groups, whereas for a dimensional construct, one locates individuals' positions along one or more latent traits. Classifying individuals into groups can be done most effectively when items maximize their discriminating power near the location of the boundary separating the groups. In contrast, locating individuals' scores along continua requires the use of some items that discriminate at each point in the full range of trait levels. Because one cannot simultaneously use items whose discriminating powers are clustered together and widely dispersed, empirically determining the latent structure of the target construct can help guide assessment.

Responding to these and related concerns, in many taxometric studies, constructs and measures of special interest to researchers and practitioners specializing in psychological assessment have been examined. Constructs studied include antisocial personality disorder (Walters & Ruscio, 2009), anxiety sensitivity (Bernstein et al., 2007), and disgust sensitivity (Olatunji & Broman-Fulks, 2007), and measures studied include the Beck Depression Inventory (BDI; A. M. Ruscio & Ruscio, 2002), Minnesota Multiphasic Personality Inventory–2 Infrequency-Psychopathology, or *F*(p), scale (Strong, Glassmire, Frederick, & Greene, 2006), and the Psychopathy Checklist: Screening Version (Walters et al., 2007). Investigators are attuned to the value of empirically assessing the latent structure of psychological constructs and measures, and methods for doing so have proliferated. In the present work, we focus on methods that are intended to differentiate categorical and dimensional data. Other methods have been developed to address related research questions involving structures that combine categorical and dimensional elements (see, e.g., Lubke & Neale, 2006; Muthén, 2001, 2006). For a discussion of using complementary data-analytic techniques to study the latent structure of psychological constructs, see J. Ruscio and Ruscio (2004).

The possibility that latent factors may represent categorical or dimensional traits has been revisited in two attempts to develop a

John Ruscio, Department of Psychology, The College of New Jersey; Glenn D. Walters, Psychology Services, Federal Correctional Institution, Schuylkill, Minersville, Pennsylvania.

The assertions and opinions contained herein are the private views of John Ruscio and Glenn D. Walters and should not be construed as official or as reflecting the views of the Federal Bureau of Prisons or the United States Department of Justice.

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

procedure for differentiating between these competing structural models with factor-analytic methods. McDonald (1967) presented an approach based on examining the distribution of estimated scores on one or more latent factors.[1] If scores on a single factor are multimodally distributed or if scores on each factor are at least bimodally distributed, one infers a categorical structural model. Otherwise, one infers a dimensional model. Steinley and McDonald (2007) tested this approach by counting the modes observed in the distributions of factor scores calculated with Bartlett's weighted least squares method (1937); a mode was defined as a local maximum in a distribution. They found that this technique identified dimensional data well but performed poorly with categorical data. This suggests either that multiple modes did not emerge for categorical data or that they did, but not in a form that met Steinley and McDonald's criterion. Locating modes in distributions of scores can be a very subjective process, and a simple, intuitively appealing criterion like the presence of a local maximum may function poorly.

To illustrate the difficulties, consider the seemingly simple case of examining the distribution of scores on the first principal factor to determine whether the data are categorical (two group) or dimensional. Unless the number of indicator variables and their validities are large, factor scores for the two groups can overlap substantially, to the point that a second local maximum does not emerge. With unequal group sizes, the distribution of factor scores for the smaller group may be swamped by the distribution for the larger group, and once again, only a single local maximum may occur in the joint distribution. On the other hand, even for dimensional data, normal sampling error can lead to a distribution with multiple local maxima. Especially if one or both tails of a distribution are long (e.g., one side of an asymmetric distribution, one or both sides of a leptokurtotic distribution), multiple local maxima may emerge. Figure 1 shows factor score distributions, in the form of density plots, for two-group categorical data sets with decreasing indicator validity (top row), categorical data in which the two groups' sizes diverge from one another (middle row), and dimensional data with increasingly skewed indicators (bottom row). If a mode is defined as a local maximum, this occurs in only three of the six distributions for categorical data, yet this condition occurs in all three of the distributions for dimensional data. It may not be possible to define a mode objectively such that more than one is identified in the categorical data plots without falsely identifying more than one in the dimensional plots. Defining a mode even more stringently may worsen the poor performance with categorical data observed by Steinley and McDonald (2007). Relaxing the criterion to define a mode without requiring a strict local maximum but rather requiring a discernible "hump" in a monotonically increasing or decreasing region of a distribution may degrade the performance with dimensional data because such apparent modes emerge rather frequently.

In their text on taxometrics, Waller and Meehl (1998), described several data-analytic procedures to test between categorical and dimensional structural models. Their latent mode (L-Mode) procedure represents a rediscovery of McDonald's (1967) method for the special case of distinguishing two-group categorical data from dimensional data (McDonald, 2003). Using L-Mode, one examines scores on the first principal factor as calculated with Bartlett's method. A bimodal factor score distribution is interpreted as evidence of categorical structure, and a unimodal distribution is interpreted as evidence of dimensional structure. In addition to testing the relative fit of these models, the L-Mode procedure takes advantage of Thurstone's (1935, 1947) observation that for categorical data, the location of the two modes can be used to estimate their relative sizes, or base rates. Thus, the L-Mode procedure also faces the challenge of identifying the number of modes in a distribution and locating each one. Waller and Meehl (1998) illustrated the potential usefulness of their L-Mode procedure to distinguish categorical and dimensional data (p. 58), and in a small-scale Monte Carlo study, Waller and Meehl found that the accuracy of the L-Mode base rate estimates compared favorably with those of cluster analyses performed with Ward's method and average linkage (pp. 61–66).

Perhaps because the problems of identifying and locating modes in factor score distributions pose substantial challenges, L-Mode has not been studied rigorously across a wide range of data conditions. In contrast, the performance of other procedures within Meehl's (1995) taxometric method has been investigated much more extensively (e.g., J. Ruscio, 2007; J. Ruscio & Marcus, 2007; J. Ruscio, Ruscio, & Meron, 2007; Walters & Ruscio, 2009). What allowed large-scale testing is a technique that objectively quantifies the graphical output of taxometric procedures with parallel analyses of categorical and dimensional comparison data (J. Ruscio et al., 2007). Rather than drawing conclusions from the results for a target data set, one generates populations of comparison data with known latent structures, draws multiple random samples from each, submits each sample of comparison data to parallel analysis, and evaluates the extent to which the results for each structure fit those of the target data. In recent years, significant advances have occurred to improve the statistical foundation and computational efficiency of each stage of this process (J. Ruscio & Kaczetow, 2008).

The primary goal in the present research was to extend this technique to the examination of factor score distributions. We adapted the approach that has been taken with other data-analytic procedures for use with L-Mode. This affords not only an objective test of the relative fit of categorical structural models versus dimensional structural models but also an estimate of the base rates.[2] All of this is accomplished by evaluating the comparative fit of factor score distributions to those for data drawn from populations with known latent structures rather than by identifying or locating modes in the factor score distributions themselves.

We evaluate this approach in a series of three studies. First, we perform an extensive Monte Carlo examination of this approach's performance across a wide range of conditions to test the robustness of L-Mode to challenging configurations of data parameters. Because the study design and random number seeds were the same as those used by J. Ruscio and Kaczetow (2009), this recreated

[1] Unless otherwise noted, all subsequent references to factor scores refer to estimated, and not true, factor scores. This terminology is used to simplify exposition, with the understanding that the procedures for examining factor-score density plots work with estimated factor scores rather than (unknown) true factor scores.

[2] Because all categorical data in the present study consisted of two groups, the groups' base rates are complementary (i.e., P and $Q = 1 - P$). From this point forward, *base rate* will refer to the proportion of cases belonging to the higher scoring group (P).

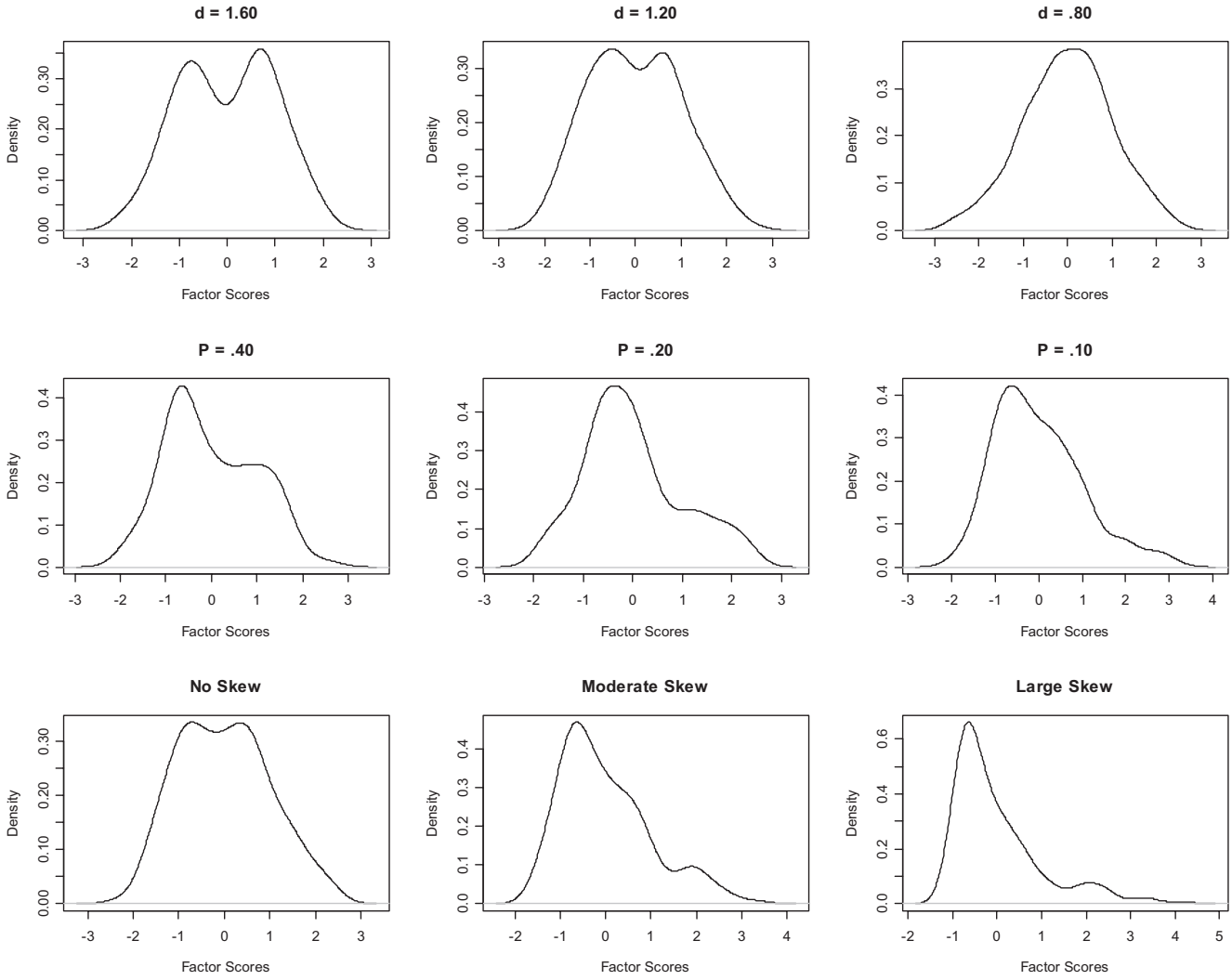

Figure 1. Top row: Factor score distributions for categorical data with two equal-sized groups but with decreasing indicator validity (validity is expressed in Cohen's *d* units). Middle row: Categorical data with large indicator validity ($d = 1.60$) but with group sizes that diverge from one another (*P* represents the proportion of the total sample in the higher-scoring group, or its base rate). Bottom row: Dimensional data with increasingly skewed indicators. Each distribution is based on a sample of $N = 200$ cases and $k = 4$ indicators.

identical target data sets and afforded comparisons of results for L-Mode with earlier findings for the maximum eigenvalue (MAXEIG; Waller & Meehl, 1998) procedure. Whereas L-Mode involves the distribution of scores on the first principal factor for all available indicators, MAXEIG involves the association among two or more output indicators within ordered subsamples of cases sorted along an input indicator. The first (largest) eigenvalue of the covariance matrix of output indicators (the variance–covariance matrix with zeros placed along the diagonal to leave only off-diagonal covariances) indexes the association among indicators within each subsample. For categorical data, output indicators should covary more strongly within subsamples approximating an equal mixture of groups than within subsamples containing mostly members of just one group. Whereas categorical data are expected to yield a peaked plot of conditional eigenvalues by subsample, dimensional data are not. Thus, MAXEIG provides a test of the relative fit of categorical and dimensional structural models that is neither conceptually nor mathematically redundant with L-Mode. A hallmark of Meehl's (1995) taxometric method is performing multiple data-analytic procedures to check the consistency of results. Whereas the output of a single data-analytic procedure may suggest a mistaken conclusion, one should be less likely to err when the results of different procedures suggest the same conclusion. In our first study, we evaluated not only the performance of L-Mode but also the use of L-Mode and MAXEIG as consistency tests for one another. As a baseline for comparison, we counted the number of modes in the distribution of scores on the first factor to evaluate the accuracy of McDonald's (1967) method as operationalized by Steinley and McDonald (2007).

In our second study, we use guidelines developed and tested in the Monte Carlo study to perform L-Mode analyses of empirical data with a known categorical structure. Meehl (1992) referred to

this as a "pseudo-problem," a way to test a procedure with real data to ensure that it provides accurate results. Finally, in our third study, we used the guidelines provided by the Monte Carlo study to perform L-Mode analyses of empirical data whose structure is presumed to be dimensional. Because dimensionality is presumed rather than known, extensive analyses are performed to check the consistency of results.

Study 1

Method

Design and Data Generation

A total of 25,000 categorical and dimensional data sets (12,500 for each structure) were generated with a Monte Carlo design in which data parameters were independently, randomly sampled from specified ranges. These data conditions were identical to those of J. Ruscio and Kaczetow (2009), which in turn were an expanded version of those of J. Ruscio (2007, Study 2) and J. Ruscio et al. (2007, Study 3). For categorical data, which consisted of two latent classes referred to as the taxon (higher-scoring group) and the complement (lower-scoring group), random values were drawn for the following parameters of each target data set: sample size (N = 300–1,000), number of indicators (k = 3–8), taxon base rate (P = .10–.50), indicator validity (standardized mean difference between classes of d = 1.25–2.00), within-group correlation (r = .00–.30), asymmetry (g = .00–.30), tail weight (h = .00–.15), and variance ratio (VR = .25–4.00; this is the ratio of variance in the taxon relative to variance in the complement).

Values of N, k, P, d, r, g, and h were drawn from uniform distributions (continuous for all but k, which was discrete) spanning the ranges listed above. The value of VR was determined by drawing a random value X from a uniform, continuous distribution ranging from 1 to 4; with probability .50, VR = X, and with probability .50, VR = 1/X. The values of g and h were used to generate data from a g-and-h distribution (Hoaglin, 1985, p. 486). The magnitude of g controls the asymmetry relative to a normal distribution (in which g = 0), and the magnitude of h controls the tail weight relative to a normal distribution (in which h = 0). Because only positive values of g and h were used, conditions of positive skew and heavy tail weight (leptokurtosis) were studied. For the g-and-h populations used in this study, smallest skew (γ_1) and kurtosis (γ_2) values were $\gamma_1 = 0$, $\gamma_2 = 0$ for $g = 0$ and $h = 0$ (a normal distribution), and the largest values were $\gamma_1 = 2.60$, $\gamma_2 = 38.89$ for $g = .30$ and $h = .15$; other pairings of g and h correspond to γ_1 and γ_2 values within this range. This covers a wide range of symmetric and asymmetric distributions that should span those encountered in most empirical data (Micceri, 1989) and pose a substantial challenge to the correct identification of latent structure.

To generate a categorical data set, the iterative technique of J. Ruscio and Kaczetow (2008) was used to sample N cases from a g-and-h distribution with $\mu = 0$, $\sigma = 1$, and a correlation matrix in which all indicators correlated r with one another. Next, a proportion P of cases was randomly selected, and those cases were identified as taxon members, with the remainder identified as members of the complement class. The desired variance ratio was achieved by multiplying scores in the taxon by a constant. Then, separation between classes was achieved by adding a constant to scores for taxon members such that the standardized mean difference equaled d.

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

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

The iterative algorithm of J. Ruscio and Kaczetow (2008) was used to sample N cases from a g-and-h distribution with $\mu = 0$, $\sigma = 1$, and a correlation matrix in which all indicators correlated r_{xy} with one another. Because VR does not correspond to a parameter of the dimensional model, it was not used in the generation of dimensional data. Extensive checking showed that our data generation programs created categorical and dimensional target data sets with the intended indicator correlations, distributions, and variance ratios. By using the same random number seeds as J. Ruscio and Kaczetow (2009) used, we recreated the same 25,000 target data sets for analysis.

Data Analysis

Counting modes. Steinley and McDonald (2007) assessed structure by counting the numbers of modes in the densities of factor score distributions. For each target data set, we performed a principal axis factor analysis restricted to a single factor and calculated factor scores with Bartlett's (1937) method. Next, a factor score density was constructed with the density function in the *R* computing environment with its default settings. This creates a density with n = 512 data points with a Gaussian kernel density estimator; by default, the bandwidth is automatically set to the standard deviation of the smoothing kernel. The number of modes in the density was counted in two ways. First, following Steinley and McDonald, we tallied the number of local maxima from the 2nd through 511th data point in the density. Categorical structure would be identified correctly by the presence of two modes, and dimensional structure would be identified correctly by the presence of one mode. Second, we trimmed 5% from each end of the density and tallied the number of local maxima within the remaining points. Trimming was done by area under the curve, not by distance along the abscissa, so unequal numbers of data points would be removed from the ends of asymmetric distributions. We expected that trimming would be helpful because sparsely populated tails of distributions may contain local maxima that represent spurious modes. We allowed only 5% trimming because some groups in the categorical data had base rates as low as .10, and trimming more of the distribution might have removed a genuine upper mode.

L-Mode. This procedure was performed by running a principal axis factor analysis restricted to a single factor and calculating factor scores with Bartlett's (1937) method. Rather than interpreting the shape of the L-Mode plot for the target data alone, parallel analyses of categorical and dimensional comparison data provided an interpretive aid.

Comparison data. To generate categorical comparison data, the base-rate classification technique was used to assign cases to

groups because prior research suggests that this approach yields good classification accuracy (J. Ruscio, 2009) and serves to differentiate categorical and dimensional target data well (see J. Ruscio, 2007; J. Ruscio & Marcus, 2007; J. Ruscio et al., 2007). Rather than estimating base rates from the results for the target data, a series of base rates was supplied to generate a series of populations of categorical comparison data. Specifically, base rates of .05, .10, .15, ...50 were provided. Each base rate was used to classify cases into the putative groups by rank-ordering cases according to their indicator total scores and applying a threshold corresponding to the proportion of the sample to be assigned to the taxon. Because no base rate is required, a single population of dimensional comparison data was generated. Within groups (for categorical data) and within the full sample (for dimensional data), a population of comparison data was generated in which indicator distributions and correlations were reproduced. L-Mode was performed for each of $B = 10$ samples drawn randomly for each of the 11 populations (10 categorical and 1 dimensional) in the same way as for the target data. The rationale for and empirical evaluation of the parallel analysis of categorical and dimensional comparison data is presented in several sources (e.g., J. Ruscio, 2007; J. Ruscio et al., 2006, 2007; Walters & Ruscio, 2009). Details of the algorithm we used to generate comparison data are provided in J. Ruscio and Kaczetow (2008),[3] and the evolution of this technique from earlier approaches is outlined in chapter 4 of the user's manual posted at www.taxometricmethod.com

It is worth noting that two very different implementations of the J. Ruscio and Kaczetow (2008) algorithm were used, one to generate samples of target data for this study and one to generate populations of categorical and dimensional comparison data to enable the calculation of an objective index to interpret the L-Mode results (see the *Comparison curve fit index* section that follows). To generate target data, the program called an outside function to generate each variable's marginal distribution from a specified *g*-and-*h* population distribution and reproduced a specified correlation matrix with a uniform intercorrelation among variables with a structural model with one latent factor. To generate comparison data, the program used a bootstrap method (resampling with replacement; see Efron & Tibshirani, 1993) to reproduce distributions of observed scores and reproduced an observed correlation matrix after determining the number of latent factors to model through a parallel analysis (Horn, 1965). The iterative algorithm for reproducing correlation matrices is the only part of the data generation procedure shared by these two approaches; aside from this, they are conceptually and statistically distinct. Moreover, both categorical and dimensional comparison data are generated with the same algorithm, so any apparent advantage provided to L-Mode by virtue of using the same data-generation technique to create target and comparison data should cancel out. The challenge is to determine whether analysis of data from the categorical or dimensional comparison population better reproduces the results for target data. We are not aware of any reason why using the same correlation-reproduction algorithm in the generation of target and comparison data would facilitate L-Mode's identification of the structure of the target data.

Comparison curve fit index (CCFI). For the target data, the distribution of factor scores was retained, and a density plot was constructed. For each population of comparison data, a single density plot was constructed with the factor scores from all random samples. CCFI values were calculated to compare the relative fit of the density plot for the target data with those of the categorical and dimensional comparison data. Because there were 10 populations of categorical comparison data, there were 10 CCFIs for each target data set (dimensional vs. categorical with base rate of .05, dimensional vs. categorical with base rate of .10, etc.). The calculation of each CCFI value had an approach adapted from the two-step process developed by J. Ruscio et al. (2007). First, the similarity between a pair of density plots was quantified with the root-mean-square residual (RMSR) of the data points. Second, the RMSRs for the dimensional and categorical comparison data were combined into a single index, the CCFI.

For data-analytic procedures such as MAXEIG, the first step in calculating the CCFI involves only the *y* values on the plots because the *x* values can be treated as equivalent across target and comparison data. For example, the *y* values on a MAXEIG plot represent eigenvalues, which are crucial to quantifying the relative fit of pairs of curves. The *x* values represent scores on the input indicator, which should be nearly identical regardless of the structure of the comparison data, because they were bootstrapped from the same univariate distribution. As expected, the CCFI has performed well when these *x* values are ignored (J. Ruscio, 2007; J. Ruscio et al., 2007). For L-Mode, however, *x* values represent factor scores in the density plot, and these can (and do) differ with the structure of the comparison data. Preliminary testing showed that a CCFI in which these *x* values were ignored did not perform well.

Our solution to this problem was to quantify the fit between a pair of density plots by calculating, for each data point on the plot for the target data, the smallest Euclidean distance to a point on the plot for the comparison data. Because the scales of the *x* and *y* axes of each density plot usually differed substantially, their ranges were equated within plots prior to quantifying fit across plots. This was accomplished by multiplying *x* values—whose mean was 0 because the factor scores had been standardized—by the ratio of the range of *y* values to the range of *x* values. Preliminary testing also revealed that factor score distributions for categorical data and

[3] An anonymous reviewer questioned whether using the same technique—the GenData algorithm of Ruscio and Kaczetow (2008)—to generate all target data sets and all populations of comparison data might inflate the performance of L-Mode. It was suggested that another technique should be used to generate target data sets, such as Waller, Underhill, and Kaiser's (1999) Monte program. This implements Vale and Maurelli's (1983) multivariate generalization of the Fleishman (1978) polynomial transformation technique for generating data with specified distributional moments. In their presentation of the GenData algorithm, Ruscio and Kaczetow (2008) discussed a number of its advantages relative to the use of polynomial transformations (e.g., a wider range of distributions can be reproduced, including those with undefined moments or different distributions sharing identical moments). Perhaps most important for the present study is that user-specified correlation matrices can be reproduced more accurately with the GenData algorithm than with the Monte program. To demonstrate this, we generated 5,000 samples of unidimensional data with a desired item-intercorrelation value chosen at random from .30 to .70. Correlations were reproduced more precisely by an implementation of the GenData algorithm than by the Monte program; the mean absolute residual correlations were .014 and .027, respectively. Due to its greater flexibility and its ability to reproduce desired correlations more precisely, we chose to use the GenData algorithm exclusively.

categorical comparison data sometimes took very similar shapes but were offset by a small amount on the x axis, perhaps because both the base rate used to generate the comparison data and the assignment of cases to groups in the comparison data were not free of error. This horizontal shift often yielded poor fit values even when a visual inspection suggested excellent fit; so, a horizontal shift parameter was estimated when the RMSR between two density plots was minimized. Thus, the first step took the following form:

$$\mathrm{Fit}_{\mathrm{RMSR}} = \sqrt{\frac{\sum (a + w_1 \times x_{\mathrm{T}} - w_2 \times x_{\mathrm{C}})^2 + \sum (y_{\mathrm{T}} - y_{\mathrm{C}})^2}{n}},$$

where x_{T} and y_{T} refer to coordinates of data points on the density plot for the target data, x_{C} and y_{C} refer to coordinates of data points on the density plot for either categorical or dimensional comparison data, w_1 and w_2 are the ratios of y to x ranges on the density plots for the target and comparison data, a is a horizontal shift parameter, and n refers to the number of points on each curve. The lowest value of this expression is calculated by allowing a to vary between -1 and $+1$; we used the optimize function in R for this purpose. Even though preliminary testing showed that a usually was estimated to be very close to 0, its inclusion improved the performance of the CCFI considerably. This fit index is calculated twice, once to quantify fit with the categorical comparison data ($\mathrm{Fit}_{\mathrm{RMSR\text{-}cat}}$) and once to assess fit with the dimensional comparison data ($\mathrm{Fit}_{\mathrm{RMSR\text{-}dim}}$). The CCFI is then calculated as

$$\mathrm{CCFI} = \frac{\mathrm{Fit}_{\mathrm{RMSR\text{-}dim}}}{\mathrm{Fit}_{\mathrm{RMSR\text{-}dim}} + \mathrm{Fit}_{\mathrm{RMSR\text{-}cat}}}.$$

CCFI values can range from 0 to 1, with lower values indicative of dimensional structure and higher values indicative of categorical structure. When $\mathrm{Fit}_{\mathrm{RMSR\text{-}dim}} = \mathrm{Fit}_{\mathrm{RMSR\text{-}cat}}$, which corresponds to equivalent fit for both structures, CCFI = .50. This represents an a priori threshold for drawing structural conclusions.

Because we obtained 10 CCFIs for each target data set, we calculated their mean for subsequent analysis. This provides a fairly conservative test of L-Mode's performance because even though some populations of categorical comparison data were generated with highly inaccurate base rates, we gave all CCFIs equal weight. For example, in Figure 2, CCFI values are plotted for a set of categorical target data with $P = .20$ by the base rate supplied to generate the 10 populations of categorical comparison data. When the base rates in the population of comparison data and the target data approached one another, the CCFI was much larger than when the base rates diverged. Nonetheless, the mean CCFI was .745, well above .50 and correctly classifying these data as categorical.

Base rate estimation. In their presentation of the L-Mode procedure, Waller and Meehl (1998) provided a technique for estimating base rates. This requires locating the latent modes that correspond to the typical factor scores for members of each group, the lower and upper modes. Because we had counted modes, we could use the Waller–Meehl (Waller and Meehl, 1998) formulas for samples that produced two (and only two) modes. This yields one estimate based on the location of the lower mode, another based on the location of the upper mode, and a third calculated as the average of these estimates. Waller and Meehl (1998, p. 60) suggested using the estimate based on the location of the upper mode when the taxon base rate is less than .50, which it was for all samples in the present study. We were able to examine the accuracy of each estimate as well as the guideline for which to prefer.

When there are not exactly two modes, one cannot use the Waller–Meehl (Waller and Mehl, 1998) approach in an objective manner. One alternative would be to subjectively identify the apparent locations of two modes, but this introduces the possibility of experimenter bias and is impractical for inspecting the many factor score distributions (for target and comparison data) that did not yield exactly two modes in the present study. Instead, we used an approach suggested by J. Ruscio (2009) that eliminates the

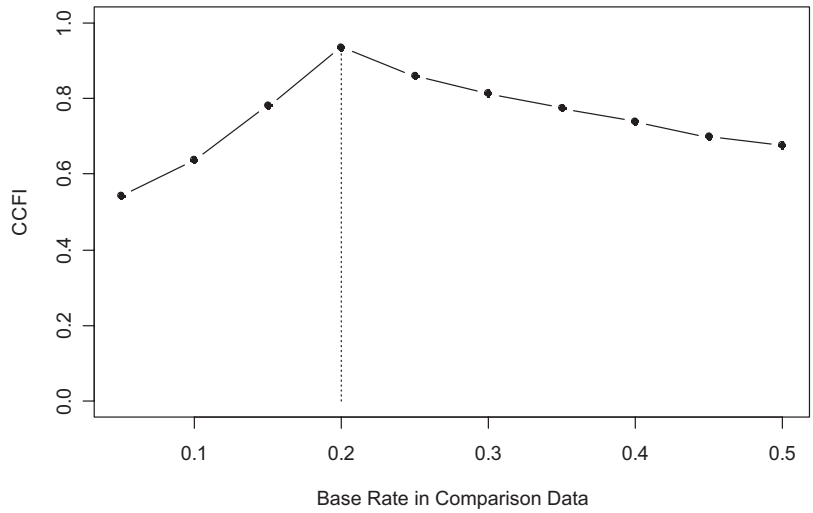

Figure 2. Comparison curve fit index (CCFI) plotted by base rate in each population of categorical comparison data. The dotted vertical line shows that the largest CCFI occurred when the population base rate was .20, very close the value of .198 in the categorical target data.

search for modes in distributions and the subjectivity of visual inspection.

Having generated multiple populations of categorical comparison data, we determined which yielded the largest CCFI and used the base rate in that population as the estimated base rate. For example, the results in Figure 2 show that the largest CCFI occurred for $P = .20$ in the population of comparison data, an estimate that comes very close to the base rate in the target data ($P = .198$). By testing the relative fit of factor score distributions to one another, the known base rates in comparison data are used to estimate the base rates in target data. This bypasses the identification and location of modes. We evaluated the potential usefulness of this technique with all 12,500 categorical target data sets.

Results

Distinguishing Categorical and Dimensional Data

The first series of analyses examined the differentiation between the 12,500 categorical and 12,500 dimensional target data sets. This was examined for the mode-counting procedure of Steinley and McDonald (2007) as well as the L-Mode procedure. The results also were compared with those for MAXEIG analyses of the same data sets (J. Ruscio & Kaczetow, 2009). Table 1 shows the numbers of modes counted for the categorical and dimensional target data sets, with the counting performed across the full density as well as the trimmed density (with the outer 5% removed from each tail). As expected, trimmed densities afforded much greater accuracy rates. For categorical data, accuracy increased from 58.6% for full densities to 74.6% for trimmed densities. For dimensional data, accuracy increased even more dramatically, from 19.5% for full densities to 94.3% for trimmed densities. With the trimmed densities, the findings are consistent with those of Steinley and McDonald (2007), in that dimensional structure was identified more accurately than categorical structure. Overall, counting the modes in trimmed densities correctly identified the latent structure of 21,108 (84.4%) of the 25,000 target data sets.

Table 1
Numbers of Modes for Target Data Sets

	Categorical		Dimensional	
Number of modes	Full density	Trimmed density	Full density	Trimmed density
1	1,276	2,531	**2,432**	**11,789**
2	**7,329**	**9,319**	4,343	706
3	2,429	592	3,220	5
4	802	55	1,645	0
5	402	3	611	0
6	174	0	170	0
7	62	0	45	0
8	22	0	8	0
9	3	0	5	0
10+	1	0	21	0
% Correct	58.6	74.6	19.5	94.3

Note. Modes were counted for the full factor score density as well as the trimmed density (middle 90%). The number of correctly identified categorical (two modes) and dimensional (one mode) data sets is indicated in bold within each column, and this is expressed as a percentage of the 12,500 data sets of each structure in the bottom row.

For L-Mode, an a priori threshold was applied such that data yielding (mean) CCFI values $> .50$ were classified as categorical, and data yielding CCFI values $< .50$ were classified as dimensional (CCFI $\neq .50$ for all target data sets). This decision rule achieved an accuracy of 98.9% for categorical data, 98.2% for dimensional data, and 98.6% overall. When these data sets were analyzed with MAXEIG, the corresponding accuracy rates for this decision rule were 90.2% for categorical data, 95.7% for dimensional data, and 93.0% overall. Each procedure's accuracy rate was considerably higher than that attained by counting modes.

In addition to testing the CCFI with a single threshold of .50, we examined the probability that the CCFI correctly identified categorical or dimensional structure across its range of values (see Figure 3). The graph on the left shows the results for L-Mode, and the graph on the right shows the results for MAXEIG. The larger the data point, the more data sets yielded CCFIs at that level. Even though MAXEIG provided strong CCFI values—data points closer to 0 or 1—more often than L-Mode, it also yielded ambiguous values—data points near .50—more often. In other words, the distribution of CCFIs for L-Mode was more strongly bimodal, with comparatively few ambiguous values and extremely few values on the wrong side of .50 (just 359 out of 25,000). Following Meehl's (2004) advice that taxometric results be sorted into those that favor categorical structure, those that favor dimensional structure, and those that are ambiguous, J. Ruscio, Walters, Marcus, and Kaczetow (2008) recommended applying dual thresholds of .45 and .55 to classify taxometric results: If CCFI $> .55$, results favor categorical structure, if CCFI $< .45$, results favor dimensional structure, and if $.45 \leq$ CCFI $\leq .55$, results are ambiguous and judgment is withheld. CCFI values fell outside the ambiguous region 96.1% of the time (24,013 out of 25,000 samples), and among these, the CCFI correctly distinguished categorical and dimensional structure 99.5% of the time (23,903 out of 24,013 samples). The corresponding figures for MAXEIG are that 91.1% of samples yielded CCFI values outside the ambiguous region, and 96.1% of these were correct.

Robustness Across Data Conditions

The fact that the accuracy of L-Mode analyses quantified with the CCFI was 98.6%—and 99.5% when ambiguous CCFI values were set aside—suggests an impressive degree of robustness across the wide range of data conditions in this study. Ceiling effects on accuracy preclude the possibility of strong effects for factors in the study design. Nonetheless, we examined the rates of accurate, ambiguous, and inaccurate results across levels of each design factor; these are displayed in Figure 4. Consistent with findings in previous studies (e.g., Beauchaine & Beauchaine, 2002; J. Ruscio & Kaczetow, 2009; J. Ruscio et al., 2007), categorical structure was more difficult to identify as indicator validity decreased or within-group correlations increased. For the other factors, trends were rather weak. More striking is that even under the most challenging levels of each factor, the error rate never exceeded 2.0%. In the few instances in which accuracy rates declined by a nontrivial amount, the results tended to be ambiguous rather than inaccurate. There were no conditions under which errors outnumbered ambiguous CCFI values. Applying dual

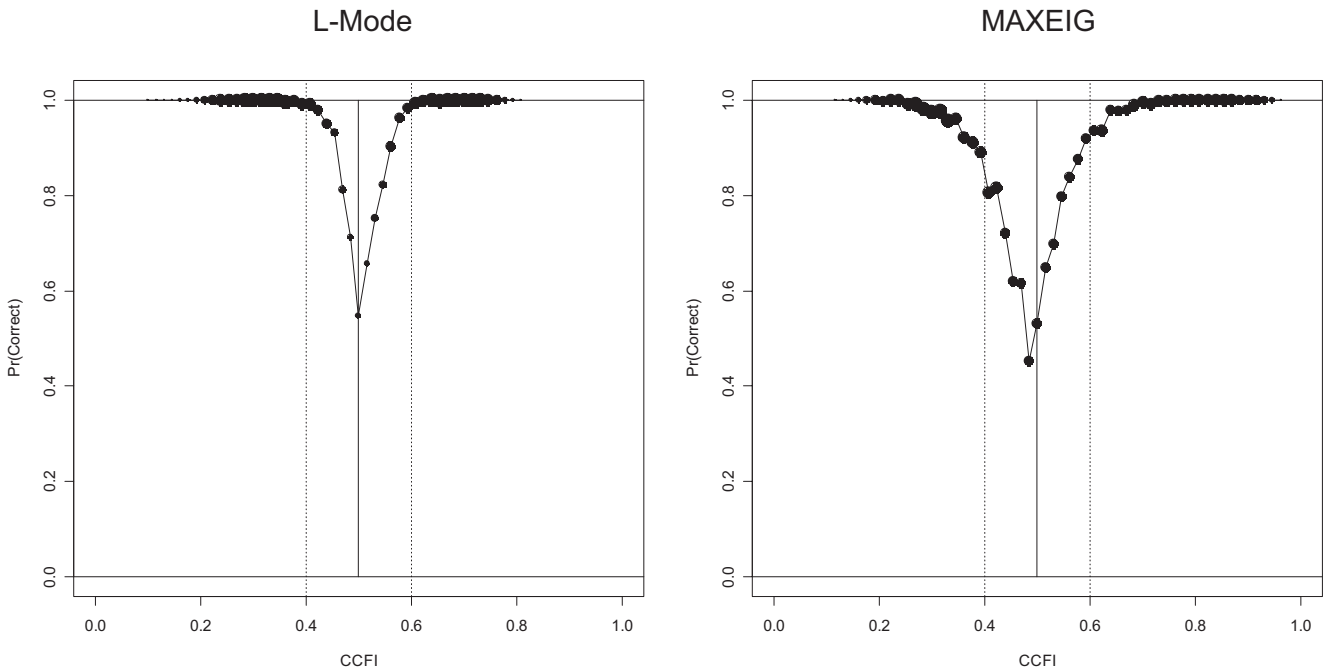

Figure 3. Probability that the comparison curve fit index (CCFI) correctly identified categorical or dimensional structure. The left graph shows the results for the present study of the latent mode (L-Mode) procedure, and the right graph shows the results of Ruscio and Kaczetow's (2009) study of the maximum eigenvalue (MAXEIG) procedure. The larger the data point, the more target data sets yielded CCFI values at that level. Vertical dotted lines provide reference points to facilitate comparisons across graphs.

thresholds of .45 and .55 to interpret CCFI values provided excellent protection against erroneous conclusions.

Though it remains possible that factors in the study design interacted with one another to predict accuracy rates, we chose not to extend this portion of the analysis of results beyond main effects. The paucity of erroneous CCFI values severely constrains the size and, therefore, the practical significance of any interaction effects. Moreover, the large number of factors that varied in the study design yields an unwieldy number of potential interaction effects, and testing even a subset of these interaction effects raises the likelihood of making Type I errors. Rather than risking the overinterpretation of weak effects, we prefer to emphasize that the most substantial risk that researchers appear to face when performing L-Mode analyses within the parameter space covered by this study is obtaining ambiguous results and not systematically misleading results. In the next section, we explore what happens when one analyzes data approaching or beyond the limits for taxometric analysis recommended by Meehl (1995).

Testing the Limits

We performed two series of tests to examine what happens as data conditions approach or exceed the limits that were built into the design of this study. First, we examined what happens when multiple factors simultaneously approach the limits. Does the CCFI offer adequate protection against mistaken conclusions? To address this question, we selected a subset of our samples that constitute the most challenging combinations of data conditions. For dimensional data, we identified 50 samples for which each of the five relevant design factors was within the most difficult third of its range (i.e., low N, k, and r_{xy}; high g and h). Among these samples, applying a single threshold of .50 to mean CCFI values achieved 92.0% accuracy, and applying dual thresholds of .45 and .55 yielded 100.0% accuracy among the nonambiguous results (42 correct, 8 ambiguous, 0 incorrect).

Too few samples would remain to perform a meaningful analysis if we selected samples according to all eight relevant factors for categorical samples. Instead, we did the same as for dimensional samples, identifying the samples that should have posed the greatest difficulty by virtue of falling within the lowest third of the range on each of the five most important factors (i.e., low N, k, P, and d; high r). The other three factors (g, h, and VR) had less influence on accuracy, so we did not select on those in order to retain enough samples to provide an informative analysis. Among the 58 samples that met these criteria, applying a single threshold yielded 82.8% accuracy, and applying dual thresholds yielded 91.3% accuracy among the nonambiguous results (42 correct, 12 ambiguous, 4 incorrect).

Pooling results across these 50 dimensional and 58 categorical samples provides 108 samples—less than one-half of 1% of the total number—that constitute the most challenging combinations of data conditions in this study. Applying dual thresholds to mean CCFI values achieved an accuracy of 95.5% among nonambiguous results (84 correct, 20 ambiguous, 4 incorrect). The CCFI provided acceptable protection against mistaken conclusions even when multiple factors were at or

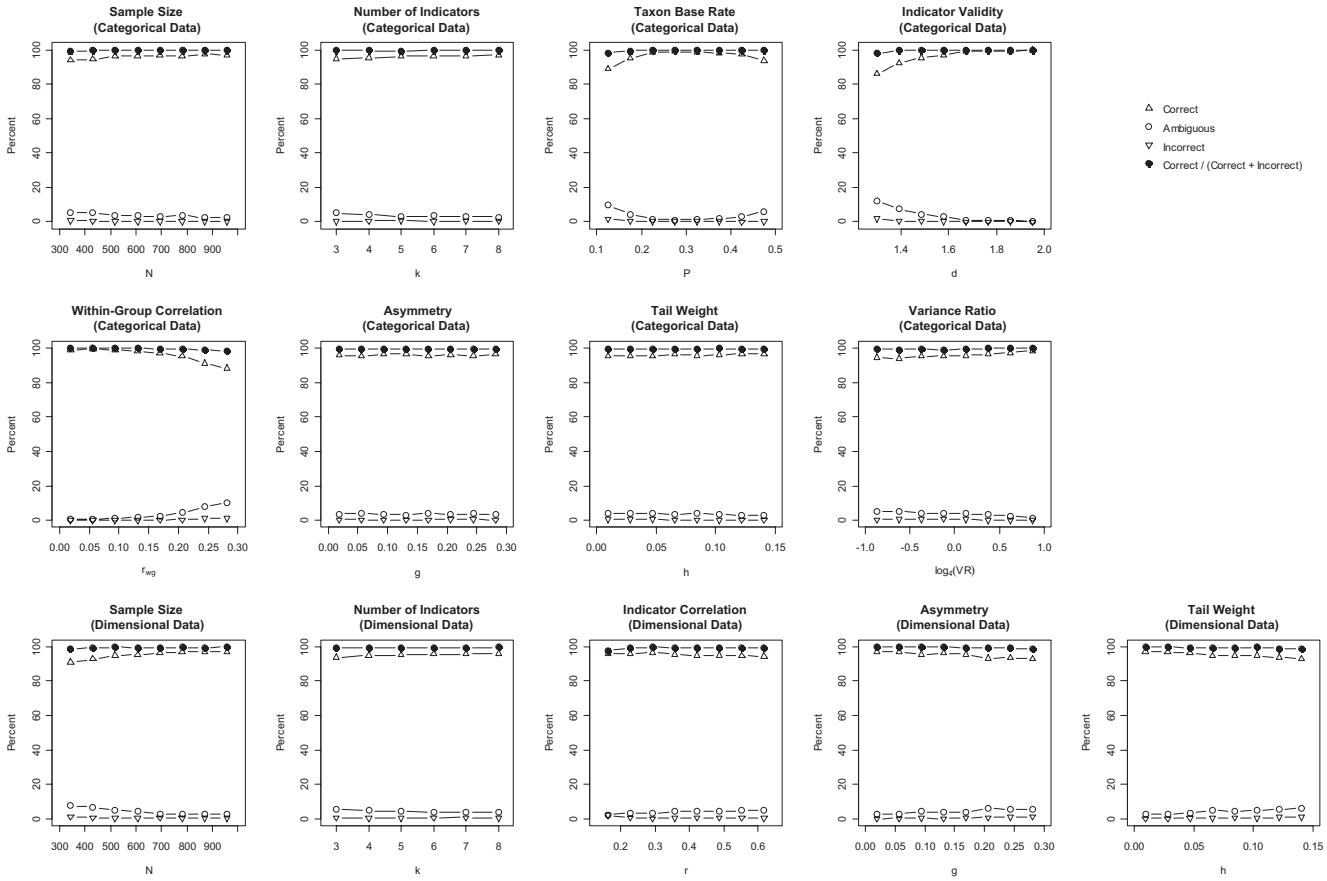

Figure 4. Accuracy rates for mean comparison curve fit index (CCFI) values by levels of each factor in the simulation study design. The percentage of incorrect values remained below 2.0% in all data conditions examined. k = number of indicators; P = taxon base rate; d = indicator validity; r_{wg} = within-group correlation; g = asymmetry; h = tail weight; VR = variance ratio.

near the recommended limits. In addition, even though the L-Mode error rate increased near the limits—as one would expect—the available alternative of counting modes produced more errors. For these 108 especially challenging samples, mode counting achieved an accuracy rate of just 20.4% with full densities (1 of 50 correct for dimensional samples, 21 of 58 correct for categorical samples) and 84% with trimmed densities (48 of 50 correct for dimensional samples, 43 of 58 correct for categorical samples). Neither of these compares favorably with the 87.0% accuracy with a single threshold for L-Mode mean CCFIs or with the 95.5% accuracy with dual thresholds.

Having examined what happens when one or more factors approaches its limits, we next tested the effects of allowing individual factors to move beyond the limits. Meehl (1995) suggested that taxometric results should be trustworthy, provided that $N \geq 300$, $P \geq .10$, $d \geq 1.25$, and $r \leq .30$. These rules of thumb were not based on published findings. Rather, they appear to reflect Meehl's (1995) experience in working with his taxometric procedures for several decades. Beauchaine and Beauchaine's (2002) study of the accuracy with which cases could be classified with the results of the maximum covariance (MAXCOV) taxometric procedure included data conditions that reached beyond the limits recommended by Meehl (1995). Performance decrements were observed, suggesting that the parameter estimates provided by this procedure degraded near Meehl's (1995) limits. Aside from this, relatively little research has involved investigating the consequences of performing taxometric analyses with data that do not satisfy Meehl's (1995) rules of thumb.

To extend the findings of the present study into this largely uncharted territory, we generated additional samples in which one factor ranged beyond values recommended by Meehl (1995) while the other factors remained in the same ranges specified earlier. Specifically, we generated categorical data with the following ranges to test each of Meehl's (1995) rules of thumb: $.05 < P < .10$ (n = 250 new samples), $.80 < d < 1.25$ (n = 1,000 new samples), $.30 < r < .50$ (n = 1,000 new samples), and $100 \leq N \leq 300$ (n = 500 new samples). Because Meehl (1995) offered no rules of thumb for dimensional data and because none of the relevant factors was associated with a precipitous decline in our results, we examined only $100 \leq N \leq 300$ (n = 500 new samples). We generated more samples when covering a wider range of new parameter values than when covering a comparatively narrow range. The results for these

3,250 new samples are presented along with those for our original samples in Figure 5, which extends the range for each factor by replotting the original beside the new results; a dotted line indicates Meehl's (1995) recommended limit. The sharp declines evident for some data parameters were real, not an artifact of data conditions or data generation. However, the appearance of declines near the recommended limits may be accentuated by grouping the original samples to plot data points within recommended limits and grouping the new samples to plot data points beyond the limits, with no samples aggregated to plot data points at the limits.

In light of these findings, Meehl's (1995) rules of thumb appear quite prescient. Beyond $P = .10$, $d = 1.25$, and $r = .30$, there was a steep decline in accuracy against which the CCFI did not afford good protection. This underscores the importance of paying careful attention to the adequacy of the data for taxometric analysis. Venturing beyond the recommended limits for P, d, or r significantly increased the risk of obtaining ambiguous or inaccurate results. At the same time, the decline in accuracy below $N = 300$ was less dramatic, and the CCFI guarded against erroneous conclusions fairly well, down to $N = 100$.

Consistency Testing

In addition to comparing the present results for L-Mode with those of previous MAXEIG analyses of the same data, we can examine the accuracy achieved when both procedures are used as consistency tests for one another. This requires that consistency be operationalized, and we did so by requiring that the CCFI for each procedure be on the same side of .50. By this standard, one would infer categorical structure if CCFI $>$ .50 for each procedure, infer dimensional structure if CCFI $<$ .50 for each procedure, and otherwise withhold judgment and reach no conclusion in the face of ambiguous results. Applied to all 25,000 target data sets, 92.5% of the results were classified as consistent, and 7.5% were classified as ambiguous. Among the consistent results, accuracy rates were 99.7% for categorical data, 99.2% for dimensional data, and 99.5% overall.

This is an excellent accuracy rate, but it is worth noting that L-Mode CCFI values achieved the same rate when dual thresholds of .45 and .55 were applied. Moreover, only 3.9% of results were set aside as ambiguous based on L-Mode CCFIs, compared with 7.5% of results when L-Mode and MAXEIG were used as consistency checks. For the added cost of discarding more results as ambiguous, consistency testing would need to achieve a greater accuracy rate than any of its constituent procedures. This did occur when consistency was operationalized more stringently by requiring that both L-Mode and MAXEIG yield CCFI values outside the ambiguous region of .45 to .55 and in the same direction (i.e., both less than .45 or both greater than .55). Among the 85.2% of samples that met this stringent criterion, accuracy was 99.8%.

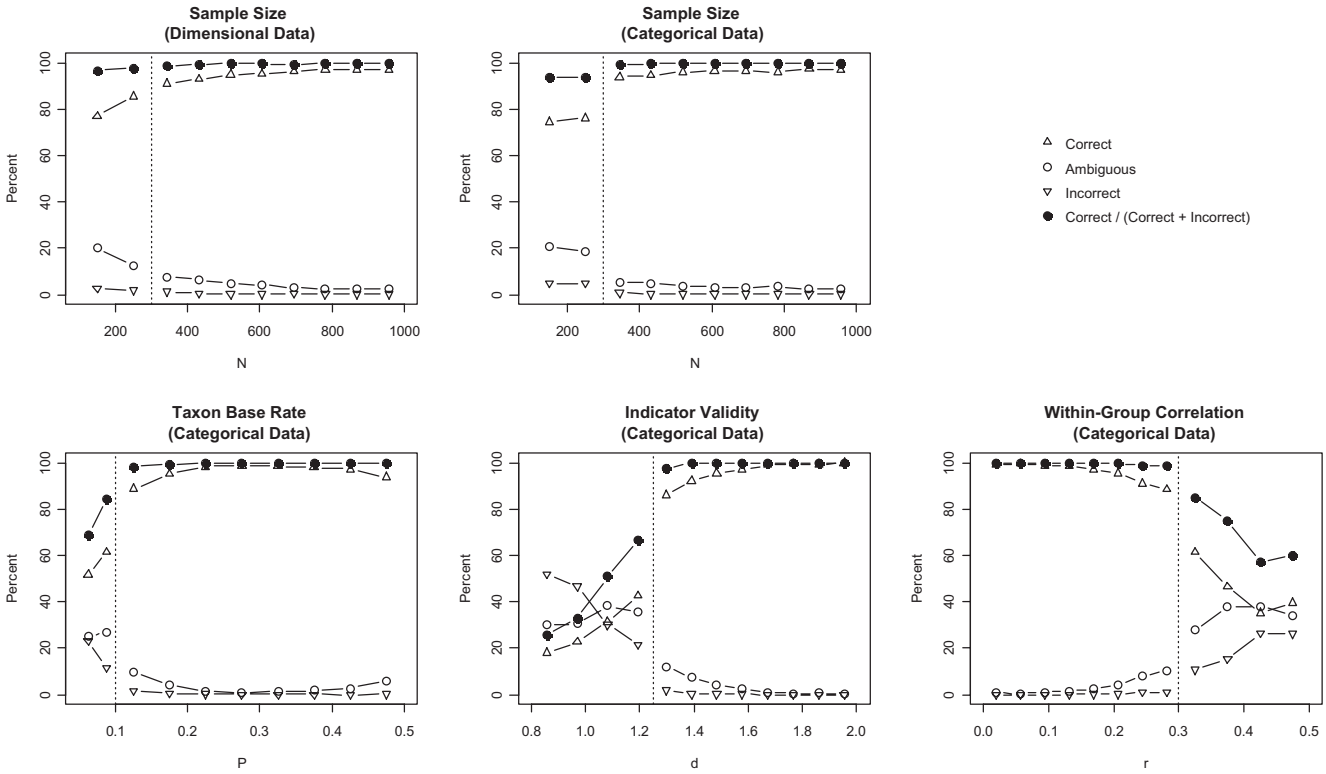

Figure 5. Accuracy rates for mean comparison curve fit index (CCFI) values by levels of factors in the simulation study design. These plots reproduce results shown in Figure 4 and extend them to wider ranges of values along each factor for examination of performance under especially challenging conditions. Dotted vertical lines represent the rules of thumb for acceptable data parameters recommended by Meehl (1995). P = taxon base rate; d = indicator validity; r = within-group correlation.

Thus, the willingness to demand stronger evidence and to withhold judgment more often is rewarded with an extremely high accuracy rate.

Accuracy of Base Rate Estimates

The final series of analyses examined the accuracy of base rate estimates. Earlier, we introduced a technique by which the base rate was estimated as the population base rate that yielded the largest of the 10 CCFI values. This means that a base rate estimate could take on a value only from the discrete list of .05, .10, .15, ...50. This constrains accuracy because actual base rates varied continuously from .10 to .50 in the target data sets. For example, if the base rate was .172 in the target data, the closest estimate possible would be .15, and this would be in error by .022. For each of the 12,500 categorical data sets, the estimated base rate was recorded. For subsets of categorical data sets that produced exactly two modes in the full densities (7,329 samples) or trimmed densities (9,319 samples), the formulas provided by Waller and Meehl (1998) were used to estimate the taxon base rate as well. For each type of estimate, its bias was calculated as the mean residual (estimated taxon base rate − actual taxon base rate) and its precision was calculated as the mean absolute residual. Results are summarized in Table 2.

For the estimates derived from comparison data, there appeared to be little or no bias (M_{bias} = .003), and precision was good ($M_{precision}$ = .028). Approximately three-quarters (74.8%) of the estimates were the closest possible value along the discrete range of possibilities, meaning that they were within ±.025 of the correct value, and most (96.9%) were no more than one value away along the discrete range of possibilities, meaning that they were within ±.075 of the correct value. J. Ruscio and Kaczetow (2009) did not record base rate estimates for the MAXEIG analyses of these data, so a comparison across procedures cannot be made.

Results for the Waller and Meehl (1998) formulas, however, provide a useful point of reference. Use of the location of the lower mode tended to an underestimation the taxon base rate (M_{bias} = −.038 and −.049 for full and trimmed densities), and use of the location of the upper mode tended to an overestimation by a larger amount (M_{bias} = .135 and .158 for full and trimmed densities); use of the mean of these estimates averages their biases (M_{bias} = .048 and .054 for full and trimmed densities). At least as important as the nontrivial bias in each of these estimates is their imprecision. Across the estimates based on lower modes, upper modes, and their means as well as full and trimmed densities, $M_{precision}$ ranged from .059 to .166. Each of these values was larger (worse) than the $M_{precision}$ for the corresponding comparison-data estimate by a factor ranging from more than 2 to nearly 6.

Finally, it is interesting to note that in terms of both bias and precision, estimates with the Waller and Meehl (1998) formula for the upper mode fared more poorly than did those for the lower mode. This is inconsistent with their recommendation to use the upper mode when $P < .50$. The failure to obtain empirical support for this recommendation may be due to the greater opportunity for sampling error to displace the upper mode within the thin tail of the density plot than the lower mode within the thicker tail, which was not considered in Waller and Meehl's analytic treatment of the problem.

Discussion

As McDonald (1967) proposed, the examination of factor score distributions appears to provide valuable clues about the latent structure that best fits a sample of data. In Steinley and McDonald's (2007) recent study, this approach did not perform very well for categorical data, a finding that occurred in the present study as well. Trimming 5% from the tails of the factor score distributions greatly improved the accuracy with which mode counting identified latent structure, but the parallel analysis of comparison data drawn from populations with known latent structures afforded still more accurate distinctions between categorical and dimensional data. Rather than identifying or locating modes, factor score distributions for target data were compared with those for comparison data, and relative fit was calculated with an objective index. Not only did this correctly identify latent structure with very high accuracy—higher than counting modes or use of the MAXEIG procedure (J. Ruscio & Kaczetow, 2009)—but it also provided fairly accurate base rate estimates. These exhibited negligible bias and good precision. In contrast, the formulas for base rate estimation introduced by Waller and Meehl (1998) require the location of two (and only two) modes, which restricts their use to those analyses that happen to yield exactly two modes, and per-

Table 2
Bias and Precision of Taxon Base Rate Estimates

	All samples		Full density		Trimmed density	
Estimate	Bias	Precision	Bias	Precision	Bias	Precision
Comparison data	.003	.028	.003	.029	.003	.027
Waller-Meehl formulas						
Lower mode			−.038	.074	−.049	.061
Upper mode			.135	.166	.158	.159
Mean			.048	.073	.054	.059

Note. For all samples, n = 12,500. For full density, n = 7,329. For trimmed density, n = 9,319. The comparison-data estimates were obtained by locating the taxon base rate from the sequence .05, .10, ...50 that when used to generate comparison data, yielded the largest comparison curve fit index value. Modes were counted for the full factor score density as well as the trimmed density (middle 90%). For samples producing exactly two modes, taxon base rate estimates were calculated with formulas provided by Waller and Meehl (1998), based on the location of the lower mode, the location of the upper mode, and the mean of these two estimates. For each estimate, bias was calculated as the mean residual (predicted taxon base rate–actual taxon base rate), and precision was calculated as the mean absolute residual.

formed poorly in terms of their bias and precision. Operationalizing consistency checks with the CCFI values provided by multiple taxometric procedures was shown to distinguish effectively between accurate results and ambiguous results that might have led to mistaken conclusions. This finding is consistent with those of a number of recent studies of taxometric consistency testing (Frazier, Ruscio, & Youngstrom, 2008; J. Ruscio, 2007; J. Ruscio et al., 2008; Walters & Ruscio, 2009).

Study 2

The simulations in Study 1 suggest that L-Mode should detect categorical structure and provide a good estimate of the base rate. In Study 2, we examine this with an approach that Meehl (1995) referred to as a "pseudo-problem," or the analysis of real data with a known structure. Specifically, we tested the performance of L-Mode with fallible indicators of biological sex. Analyses of empirical and comparison data were performed, and results were interpreted, following the guidelines provided in Study 1.

Method

Data were drawn from the Hathaway Data Bank, which contains responses to the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) for a very large sample of individuals at hospitals affiliated with the University of Minnesota from 1940 to 1976. Following the procedures described in J. Ruscio and Ruscio (2000, Study 2), we identified a single, valid record for each patient in the database. The 40 items on the Masculinity–Femininity scale (Scale 5, *Mf*) that best differentiated men and women were assigned to three indicators. The first 14 items (1, 4, 7, 38, 56, 70, 74, 77, 78, 81, 87, 92, 118, 132) formed Indicator 1, the second 13 items (140, 144, 149, 158, 176, 215, 219, 223, 261, 283, 294, 295, 300) formed Indicator 2, and the remaining 13 items (367, 392, 427, 434, 463, 522, 537, 538, 539, 550, 557, 563, 566) formed Indicator 3. Complete data were available for $N = 13{,}815$ cases, including 5,671 men (41.0%) and 8,144 women (59.0%). Items were keyed such that the smaller group (men) scored higher. Indicators were mildly positively skewed in the full sample (M skew = .16), negatively skewed among men ($M = -.50$), and positively skewed among women ($M = .62$). Indicator validities ranged from $d = 1.31$ to 1.68 ($M = 1.49$), and within-group correlations were substantial among men ($M = .40$) and women ($M = .36$). These correlations fall beyond the limits recommended by Meehl (1995), and in Study 1 we found that this increases the risk that categorical data will yield inaccurate results. If we were to obtain apparently dimensional results, it would be difficult to draw a structural conclusion with much confidence. On the other hand, if we were to obtain apparently categorical results despite the large within-group correlations, it would be reasonable to reach a conclusion of categorical structure. Because this was an unusually large sample, we selected 500 cases at random for follow-up analysis, and this subset included 208 men and 292 women.

Results and Discussion

L-Mode was performed as in Study 1, with a single population of dimensional comparison data and 10 populations of categorical comparison data with base rates ranging from .05 to .50. Because it was feasible in Study 2, we analyzed 100 samples of comparison data—rather than 10, as in Study 1—for each population. For the full sample and the smaller subset, mean CCFI values of .651 and .634, respectively, correctly identified the categorical structure of these data; this occurred despite the large within-group correlations that would have been cause for concern had the CCFI values suggested dimensional structure. The results for the subset of $n = 500$ cases are shown in Figure 6. Though there was a discernible hump on the right side of the curve, there was only one local maximum (at a factor score of −.69). Thus, the mode-counting approach would fail to identify the categorical structure of these data, suggesting dimensional structure instead (the same was true for the full sample of $N = 13{,}815$). Not only was L-Mode able to identify categorical structure, but the largest CCFI was observed for the population of categorical comparison data generated with a base rate of .40, close to the actual base rate of men in this sample (208/500 = .416). The same pattern emerged in the full-sample analysis, with the largest CCFI at a base rate of .40, close to the actual base rate of .410. These results are consistent with those of the simulations in Study 1 in showing that L-Mode can correctly identify categorical structure, even when counting local maxima does not, as well as provide a good estimate of the base rate.

Study 3

In our next study, we examined the mode-counting and L-Mode techniques as applied to a Big Five personality construct whose structure is widely presumed to be dimensional (see Haslam & Kim, 2002, for a discussion of the structure of broad vs. narrow traits and a review of pertinent evidence). Because the dimensional structure of personality traits is not as firmly established as the categorical structure of biological sex, we performed more extensive analyses to check the consistency of findings.

Method

Data were drawn from the same large database of MMPI responses as in Study 2. This time, we analyzed the 70 items on the Social Introversion scale (Scale 10, *Si*). Complete data were available for $N = 13{,}581$ cases. Rather than assigning one third of the items to each of three indicators, we randomly assigned items to $k = 3$, 4, 5, 6, or 7 indicators. This was done with the full sample of data as well as subsets of $n = 650$ randomly selected cases (the midpoint of the range used in Study 1). Each combination of sample size and number of indicators was fleshed out with 10 replications with different random item parcels. For each of these 100 analyses, we calculated the mean CCFI and counted the number of modes in the full and trimmed factor score densities. This afforded an examination of the consistency of results for each method.

Because each of these 100 analyses involves a different set of indicators (and when $n = 650$, a different random subset of cases was used for each analysis) we cannot provide estimates of data parameters for all configurations. Instead, we offer an overview based on the full sample of data. There was only a modest amount of indicator skew (mean skew = .25), and indicators were strongly correlated with one another (mean $r = .65$). When the highest-scoring 50% of cases was assigned to one group and the lowest-scoring 50% to the other group, the sequential assignment of items to three indicators yielded strong between-group validities (mean

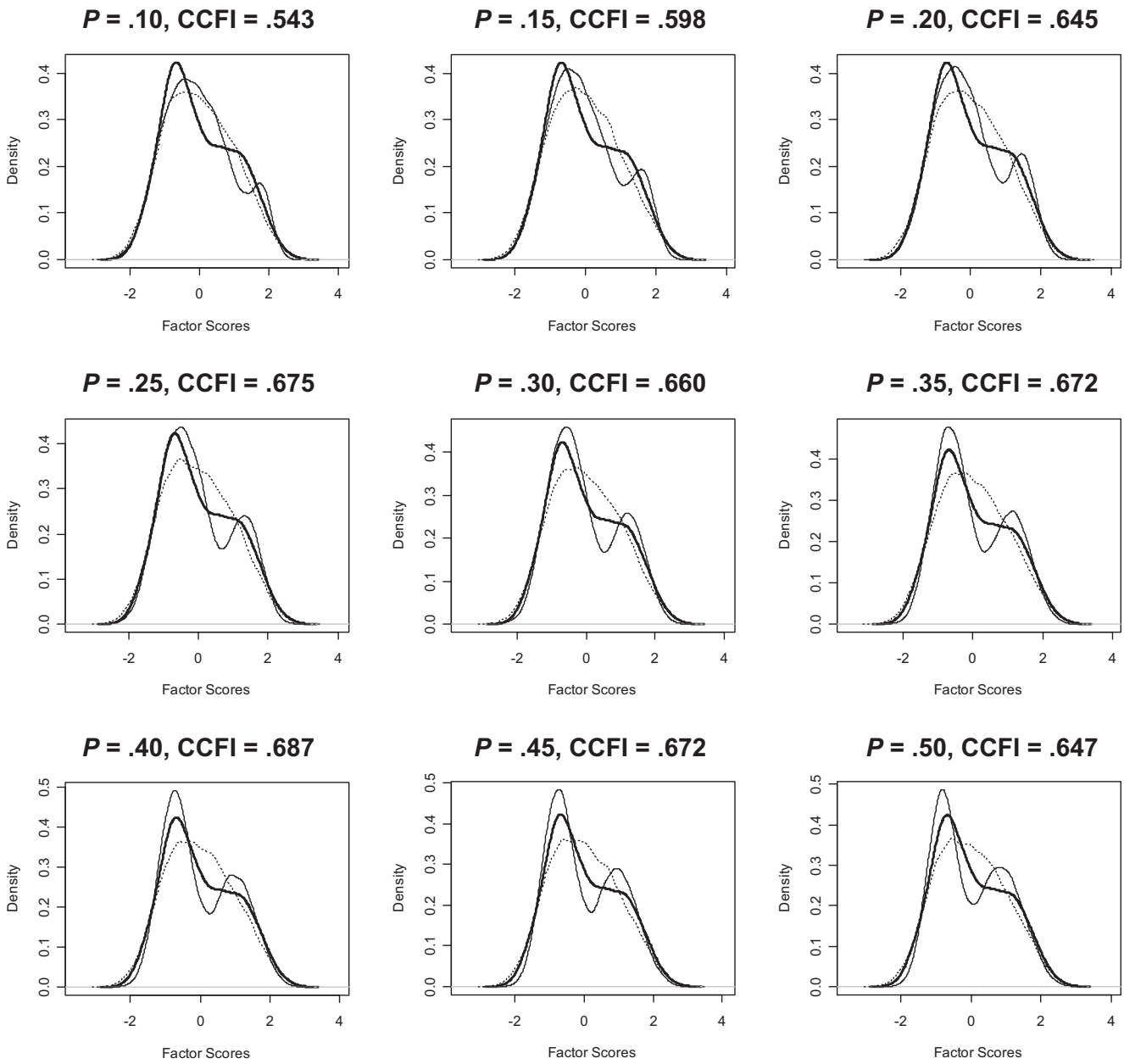

Figure 6. Latent mode (L-Mode) graphs for analyses of the Minnesota Multiphasic Personality Inventory (MMPI) Masculinity–Femininity data with $n = 500$. Each graph contains the density plots for the MMPI data (dark solid line), categorical comparison data (lighter solid line), and dimensional comparison data (dotted line). Graphs are labeled with the base rate used to generate the categorical comparison data (P) and the comparison curve fit index (CCFI) observed for that analysis. The graph for $P = .05$, CCFI = .545 is not shown, and the mean CCFI across all 10 analyses was .634.

$d = 2.08$) and nontrivial within-group correlations (mean $r = .32$ in the higher-scoring group, mean $r = .21$ in the lower-scoring group). Thus, the data appear adequate for taxometric analysis in that categories could be detected if they existed.

Results and Discussion

The mode-counting approaches performed poorly with these data. For example, the factor score density shown at the top of Figure 7 emerged when the 70 items were assigned to 5 indicators containing 14 items apiece. The apparent modes most likely arise as an artifact of the MMPI's dichotomous item response format. Depending on factors such as the number of items, the item response scales, the number of composite indicators created from the items, and the loadings of items or composites on the common factor used to estimate factor scores, the density of the factor scores may or may not take on the spiked appearance seen in this graph. For example, the

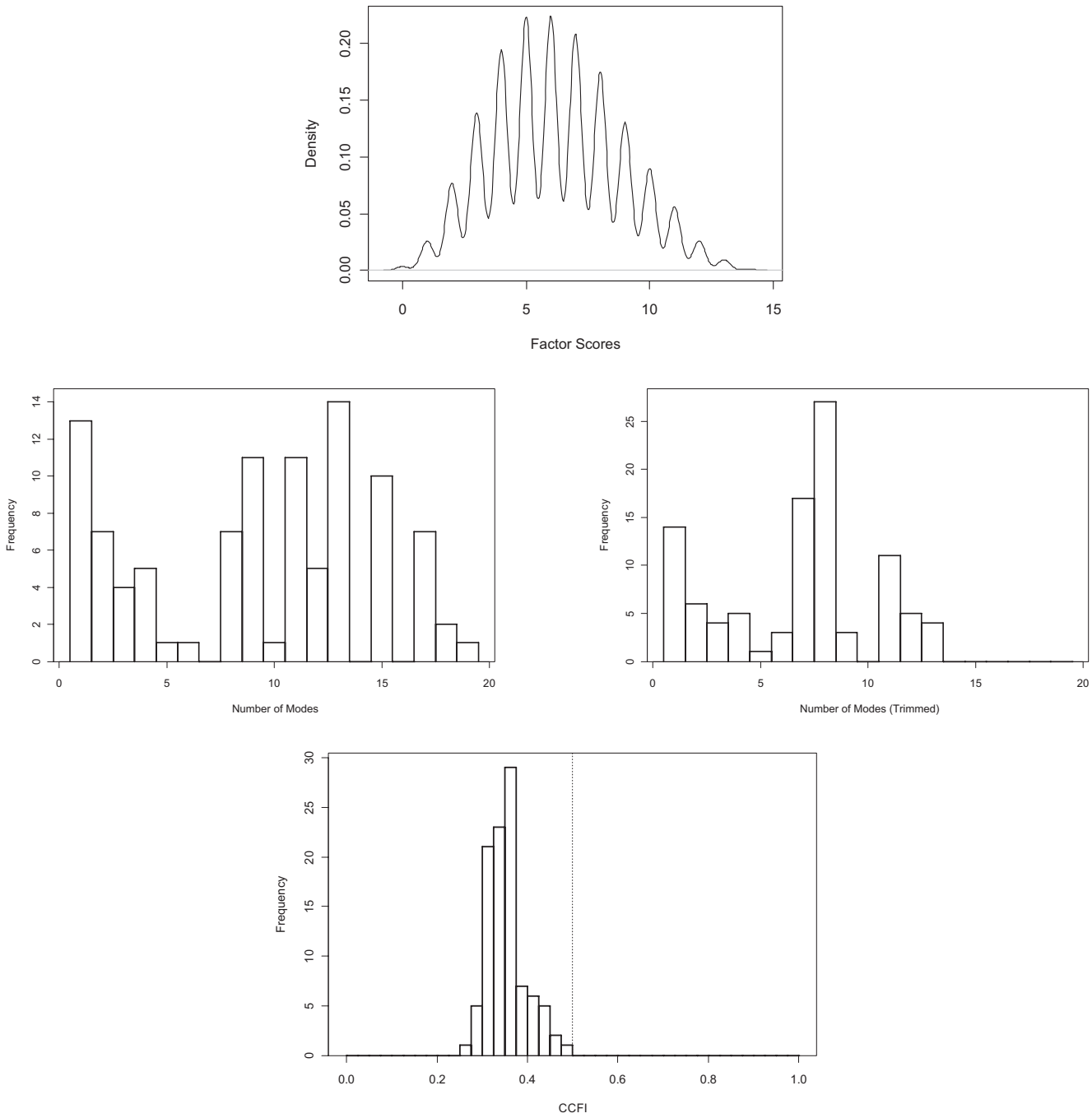

Figure 7. Results for analyses of the Minnesota Multiphasic Personality Inventory (MMPI) social introversion data. The top graph shows an illustrative factor score density plot for 5 indicators created by randomly summing 14 items apiece. There are 15 local maxima (modes), 8 after trimming the distribution. The comparison curve fit index (CCFI) for an L-Mode analysis of those data yielded CCFI = .366. The pair of histograms in the middle row shows the counts of modes in factor score densities (full densities at left, trimmed densities at right), and the bottom histogram shows all CCFI values.

factor score density for the MMPI Masculinity–Femininity items in Study 2 was not spiked. The fact that this effect can occur, however, represents a significant limitation of the mode-counting approach to studying latent structure via factor score densities because psychological assessment and research often involves dichotomous items or ordered categorical data such as parcels of dichotomous items. Across the 100 analyses of MMPI social introversion data, the number of modes was

highly variable, ranging from 1 to 19 for full densities and from 1 to 13 for trimmed densities (see histograms in Figure 7). Presuming that social introversion is not a categorical construct, the correct value of 1 mode occurred for only 13% of the full densities and 14% of the trimmed densities.

In contrast, all CCFI values obtained through L-Mode analyses with parallel analyses of categorical and dimensional comparison data favored dimensional structure (i.e., all CCFIs $<$ .50; see Figure 7). Applying dual thresholds at .45 and .55, 94% of the CCFIs favored dimensional structure (i.e., CCFI $<$ .45), and the remaining 6% were in the intermediate range of ambiguous values. Unlike the counted modes, the L-Mode results were consistent across large and small samples with items randomly assigned to as few as three or as many as seven indicators. These results also are consistent with the presumption (and evidence; see Haslam & Kim, 2002) that social introversion is a dimensional personality construct.

General Discussion

The results of each of these three studies provide support for parallel analyses of categorical and dimensional comparison data to help interpret factor score density plots. Whereas counting modes, defined as local maxima, correctly identified structure for 84.4% of the 25,000 target data sets in Study 1, the L-Mode procedure achieved an accuracy rate of 98.6% when the CCFI was calculated as an objective index of the relative fit of categorical and dimensional structural models. When dual thresholds were applied to establish an intermediate range of ambiguous results, the accuracy rate among the 96.1% of samples that passed this screen was 99.5%. L-Mode analyses, quantified objectively with the CCFI, led to relatively few mistakes, and approximately two thirds of the errors were avoided by setting aside a small percentage of comparatively ambiguous results. Provided that each remained within the recommended limits for taxometric analysis, even when multiple factors approached the limits, the CCFI provided good protection against mistaken conclusions. It was not until factors crossed these limits that inaccurate results increased markedly, and the CCFI failed to provide adequate protection. The inclusion of comparison data in L-Mode analyses afforded estimates of the taxon base rate that were unbiased and precise, in contrast to the base rate estimates that were calculated from the locations of modes.

In Study 2, mode counting failed to identify a structure known to be categorical (biological sex). L-Mode not only identified the categorical structure of these data but also provided a good estimate of the taxon base rate. In Study 3, mode counting yielded quite discrepant results across analyses of the same data and was susceptible to the identification of spurious categories—sometimes a large number of them. Though this was almost certainly an artifact of the dichotomous response format of the items used to construct indicators for analysis, psychological assessment instruments often contain items or yield composite scores that vary across a relatively small number of discrete values. This suggests that one should count the modes in factor score density plots with caution unless the data approximate continuous scales. In contrast, L-Mode yielded results consistent with one another and with the presumption that broad personality traits are dimensional constructs (Haslam & Kim, 2002).

In each of these three studies, populations of categorical comparison data were generated with a very broad range of base rates (.05 to .50). This was done to provide a conservative test of the approach by making the procedure blind to an a priori base rate estimate. By giving equal weight to results based on categorical comparison data generated with both accurate and inaccurate base rates, this may have weakened the performance of the average CCFI value. In actual investigations, researchers can take advantage of hypothesized base rates to generate more appropriate populations of categorical comparison data. On the other hand, averaging multiple CCFI values should have reduced the standard error of the composite relative to that of individual CCFI values. Use of a composite CCFI value might improve the accuracy with which categorical and dimensional data are differentiated. Perhaps the best strategy would be to use a more narrow range of plausible base rate estimates to generate populations of categorical comparison data, calculating and interpreting the mean CCFI. For example, if one hypothesizes a base rate of $P = .25$, populations of comparison data can be generated with taxon base rates of .20, .21, .22, ...30. This would yield 11 CCFI values to average, each based on a plausible base rate estimate.

Data from Study 1 illustrate the potential value of narrowing the range of base rate estimates used to generate populations of categorical comparison data. The mean CCFI was recalculated with the three values for the base rates closest to that for each categorical target data set. For example, if the actual base rate was .172, the three closest base rates in the comparison data were .10, .15, and .20, and the resulting CCFIs for these three populations of comparison data would be averaged. Whereas use of the mean of all 10 CCFIs correctly identified 12,360 (98.9%) of the categorical samples, use of the mean of 3 CCFIs at the closest base rates correctly identified 12,458 (99.7%) of the categorical samples; put another way, the latter reduced errors by 70% (from 140 errors to 42 errors).

A more carefully chosen array of base rates might yield more accurate base rate estimates as well. In the example of a hypothesized taxon base rate of $P = .25$, use of base rates of .20 through .30 in increments of .01 provides the opportunity to estimate the base rate more precisely than in the present study, in which increments of .05 were used between successive base rates. With a sufficiently numerous array of closely spaced base rates, one could smooth a plot of CCFIs by base rate to locate the maximum CCFI value and use the corresponding base rate as the best estimate of its value in the target data. In future research, the usefulness a range of more carefully targeted base rates used to generate populations of categorical comparison data should be investigated.

The examination of factor score distributions via the L-Mode procedure surpassed the impressive accuracy with which the MAXEIG procedure distinguished categorical and dimensional data in a previous study with the same 25,000 target data sets as in Study 1. Moreover, these procedures served well as consistency tests for one another. Meehl (2004) took every opportunity to emphasize the importance of checking the consistency of results. In his last article on the taxometric method, he wrote the following:

> I have always advocated that taxometricians should use multiple taxometric procedures and consistency tests, and I have called my taxometric method coherent cut kinetics to emphasize that the results will be in reasonable agreement if the underlying situation is a certain structure. If the latent structure is [categorical[4]], one sort of coherent picture will emerge; if it is [dimensional], a different sort of picture will emerge; if what emerges is unclear, judgment should be suspended until more evidence is examined. (Meehl, 2004, p. 42)

Though the central message of the importance of consistency checks resonates throughout the literature on his taxometric method, until recently it had not been operationalized in a way that is both objective and empirically supported. The present results add to a growing body of evidence in support of performing parallel analyses of comparison data and calculating CCFI values for multiple taxometric procedures, and then requiring that these CCFI values meet a specified criterion for consistency (Frazier et al., 2008; J. Ruscio, 2007; J. Ruscio et al., 2008; Walters & Ruscio, 2009). This approach allows one to sort results into the three piles that Meehl (2004) discusses: consistent evidence of categorical structure, consistent evidence of dimensional structure, and ambiguous evidence that suggests withholding judgment until additional evidence clarifies matters. J. Ruscio et al. (2008) examined taxometric consistency testing with the broadest range of data-analytic procedures and provided the most rigorously evaluated guidelines for operationalizing consistency testing.

In addition to facilitating the operationalization of consistency testing, the present research provides strong empirical support for Meehl's (1995) recommended limits to the data conditions amenable to taxometric analysis. Specifically, the putative taxon and complement should meet the following criteria: taxon base rate of $P \geq .10$, indicator validity of $d \geq 1.25$, and within-group correlations of $r \leq .30$. Beyond these values, the accuracy of results declined steeply, and the CCFI did not offer good protection against mistaken conclusions. Meehl's (1995) recommendation that N be at least 300 was supported more weakly, as the dropoff in accuracy was more modest than for the other factors, and the CCFI offered reasonable protection down to $N = 100$. Even though accuracy decreased sharply beyond the limits suggested by Meehl (1995), when multiple factors approach the limits but remain within the acceptable ranges, it appears that one's greatest risk is obtaining ambiguous—but not inaccurate—results.

The present findings provide even stronger support for an approach to the study of categorical and dimensional latent structures pioneered by McDonald (1967) than does the recent work of Steinley and McDonald (2007). By comparing distributions of factor scores to those for comparison data drawn from populations with known latent structures rather than attempting to identify modes, both categorical and dimensional data were identified more accurately than by counting the number of modes. This structural distinction was achieved successfully across a wide range of data conditions. Whereas we focused on Waller and Meehl's (1998) L-Mode procedure, which is used to attempt to differentiate categorical from dimensional data in the special case of only two hypothesized groups by examination of the distribution of scores on the first principal factor, perhaps it would be possible to extend our approach to the examination of multiple factor score distributions to compare the relative fit of more complex structural models. Bartholomew (1987) showed that correlation matrices can be reproduced equally well with a structural model with k latent factors (dimensions) or $k + 1$ latent classes. McDonald's (1967) method allows one to compare the relative fit of latent structures with $k > 1$. Even though Steinley and McDonald's initial results were discouraging for the identification of categorical data, further study is warranted to determine whether the parallel analysis of comparison data might achieve better results. Specifically, one can use the data-generation algorithm of J. Ruscio and Kaczetow (2008) to reproduce the observed distributions and correlations in a multivariate nonnormal set of psychological assessment data with various known latent structures. Populations of one-dimensional, two-dimensional, . . . k-dimensional and one-group, two-group, . . . $(k + 1)$-group comparison data can be generated and submitted to parallel analysis for the purpose of determining which set of factor-score density plots best reproduces those of an empirical data set.

Parallel analysis of appropriately generated comparison data has proven successful in the realms of factor analysis (Fabrigar, Wegener, MacCallum, & Strahan, 1999) and taxometric analysis (J. Ruscio et al., 2007). This facilitates research involving the examination of the number of factors underlying a construct or the relative fit of a two-group categorical model and a purely dimensional model. Investigators may wish to test the fit of more complex structural models that represent theoretically or practically important distinctions. For example, one might want to evaluate the relative fit of a model representing individuals who meet criteria for a mental disorder, versus those who do not meet criteria for a mental disorder, versus a three-group model that also includes an intermediate, subthreshold, group. Whereas Waller and Meehl's (1998) L-Mode procedure was not designed to differentiate between two-group and three-group categorical models, McDonald's (1967) method could be used. Counting modes in the density plots for multiple factors poses the same challenge as counting modes in a single plot, but parallel analyses of appropriately generated comparison data may overcome this obstacle. Rather than counting or identifying modes in each factor score distribution, the relative fit of comparison data representing each proposed structural model can be quantified to infer the structure of the data. Theory and practice in assessment raises a number of questions about the categorical and dimensional structures of psychological constructs. The parallel analysis of comparison data may prove useful for many data-analytic approaches to addressing these questions.

[4] In his articles, chapters, and book on the taxometric method, Meehl consistently stated or implied that the structural distinction is between taxonic (categorical) and nontaxonic (dimensional). For example, the subtitle of his coauthored book on taxometrics (Waller & Meehl, 1998) is *Distinguishing Types From Continua*, that is, the subject is categories versus dimensions. In every illustrative nontaxonic data set analyzed in this book—as well as the nontaxonic data sets analyzed in the monographs of Meehl and Yonce (1994, 1996)—the latent structure is dimensional.

References

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

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28,* 97–104.

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.

Bernstein, A., Zvolensky, M. J., Norton, P. J., Schmidt, N. B., Taylor, S., Forsyth, J. P., et al. (2007). Taxometric and factor analytic models of anxiety sensitivity: Integrating approaches to latent structural research. *Psychological Assessment, 19,* 74–87.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

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

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521–532.

Frazier, T. W. Ruscio, J., & Youngstrom, E. A. (2008*). Comparing taxometric analysis and latent variable models in the modeling of categorical and dimensional structure*. Manuscript in preparation.

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

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

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory* (Rev. ed.). Minneapolis, MN: University of Minnesota Press.

Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*-and-*h* distributions. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data, tables, trends, and shapes* (pp. 461–511). New York: Wiley.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Lubke, G., & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499–532.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monograph, 15, 32*(4, Pt. 2).

McDonald, R. P. (2003). A review of multivariate taxometric procedures: Distinguishing types from continua. *Journal of Educational and Behavioral Statistics, 28,* 77–81.

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,* 266–275.

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.

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: American Psychological Association.

Muthén, B. (2006). Should substance use disorders by considered as categorical or dimensional? *Addiction, 101*(Suppl. 1), 6–16.

Olatunji, B. O., & Broman-Fulks, J. J. (2007). A taxometric study of the latent structure of disgust sensitivity: Converging evidence for dimensionality. *Psychological Assessment, 19,* 437–448.

Ruscio, A. M., & Ruscio, J. (2002). The latent structure of analogue depression: Should the BDI be used to classify groups? *Psychological Assessment, 14,* 135–145.

Ruscio, J. (2007). Taxometric analysis: An empirically grounded approach to implementing the model. *Criminal Justice and Behavior, 34,* 1588–1622.

Ruscio, J. (2009). Assigning cases to groups using taxometric results: An empirical comparison of classification techniques. *Assessment, 16,* 55–70.

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method: A practical guide.* Mahwah, NJ: Erlbaum.

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43,* 355–381.

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the robustness of taxometric analysis. *Multivariate Behavioral Research, 44,* 259–280.

Ruscio, J., & Marcus, D. K. (2007). Detecting small taxa using simulated comparison data: A reanalysis of Beach, Amir, and Bau's (2005) data. *Psychological Assessment, 19,* 241–246.

Ruscio, J., & Ruscio, A. M. (2000). Informing the continuity controversy: A taxometric analysis of depression. *Journal of Abnormal Psychology, 109,* 473–487.

Ruscio, J., & Ruscio, A. M. (2004). Clarifying boundary issues in psychopathology: The role of taxometrics in a comprehensive program of structural research. *Journal of Abnormal Psychology, 113,* 24–38.

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research, 42,* 349–386.

Ruscio, J., Walters, G. D., Marcus, D. K., & Kaczetow, W. (2008). *Comparing the relative fit of categorical and dimensional latent variable models using consistency tests*. Manuscript submitted for publication.

Steinley, D., & McDonald, R. P. (2007). Examining factor score distributions to determine the nature of latent spaces. *Multivariate Behavioral Research, 42,* 133–156.

Strong, D. R., Glassmire, D. M., Frederick, R. I., & Greene, R. L. (2006). Evaluating the latent structure of the MMPI-2 F(p) scale in a forensic sample: A taxometric analysis. *Psychological Assessment, 18,* 250–261.

Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Vale, D. C., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465–471.

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 simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. *Multivariate Behavioral Research, 34,* 123–142.

Walters, G. D., Gray, N. S., Jackson, R. L., Sewell, K. W., Rogers, R., Taylor, J., & Snowden, R. J. (2007). A taxometric analysis of the Psychopathy Checklist: Screening Version (PCL:SV): Further evidence of dimensionality. *Psychological Assessment, 19,* 330–339.

Walters, G. D., & Ruscio, J. (2009). To sum or not to sum: Taxometric analysis with ordered categorical assessment items. *Psychological Assessment, 21,* 99–111.

Received November 19, 2008
Revision received April 20, 2009
Accepted May 12, 2009 ■