

# Comparing the Relative Fit of Categorical and Dimensional Latent Variable Models Using Consistency Tests

John Ruscio  
The College of New Jersey

Glenn D. Walters  
Federal Correctional Institution, Schuylkill

David K. Marcus  
University of Southern Mississippi

Walter Kacetow  
Graduate Center of the City University of New York

A number of recent studies have used Meehl's (1995) taxometric method to determine empirically whether one should model assessment-related constructs as categories or dimensions. The taxometric method includes multiple data-analytic procedures designed to check the consistency of results. The goal is to differentiate between strong evidence of categorical structure, strong evidence of dimensional structure, and ambiguous evidence that suggests withholding judgment. Many taxometric consistency tests have been proposed, but their use has not been operationalized and studied rigorously. What tests should be performed, how should results be combined, and what thresholds should be applied? We present an approach to consistency testing that builds on prior work demonstrating that parallel analyses of categorical and dimensional comparison data provide an accurate index of the relative fit of competing structural models. Using a large simulation study spanning a wide range of data conditions, we examine many critical elements of this approach. The results provide empirical support for what marks the first rigorous operationalization of consistency testing. We discuss and empirically illustrate guidelines for implementing this approach and suggest avenues for future research to extend the practice of consistency testing to other techniques for modeling latent variables in the realm of psychological assessment.

*Keywords:* latent variable modeling, consistency testing, categories, dimensions, taxometric method

Methodologists have developed a number of data-analytic techniques for determining whether a categorical (latent class, qualitative, taxonic) or dimensional (latent factor, quantitative, continuous) model better fits a set of data. Empirically evaluating the structure of psychological constructs can have implications for assessment, classification, diagnosis, research design, statistical analysis, and theorization (MacCallum, Zhang, Preacher, & Rucker, 2002; Meehl, 1992; Ruscio, Haslam, & Ruscio, 2006). For instance, when validating a measure that assesses a categorical construct the focus would be on classification accuracy, whereas when validating a measure that assesses a dimensional construct the focus would be on correlations with associated features. The umbrella of construct validation might include whether the scores on a measure match the latent structure of the construct (e.g., accurate cut scores are essential for assessing a

categorical construct, but a measure of a dimensional construct should not yield dichotomous scores).

Techniques to differentiate categorical and dimensional data include cluster analysis (Everitt, Landau, & Leese, 2001), mixture modeling (McLachlan & Peel, 2001), latent class or latent profile analysis (McCutcheon, 1987), and Meehl's (1995) taxometric method. When performing any of these techniques, researchers are faced with a number of choices that must be made to implement the analyses for a particular empirical data set. Ordinarily, multiple models are specified and the best-fitting model is retained. Whereas replications in new samples may be performed to test the relative fit of the same models using the same data-analytic procedures, it would be unusual to perform a wide range of complementary procedures to examine the coherence of results for a single sample of data.

Meehl's (1995) taxometric method constitutes an exception to this rule because it includes multiple data-analytic procedures expressly designed to afford checks on the consistency of results. Each taxometric procedure provides one or more graphs as well as estimates of structural model parameters. These results can be inspected to determine whether they consistently support a categorical or a dimensional model, or whether the results are ambiguous and neither structural model is better supported relative to the other. Meehl and his colleagues (e.g., Grove, 2004; Meehl, 1995; Meehl & Golden, 1982; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998) emphasized consistency testing as a cornerstone of this method, a necessary component of a taxometric investigation that should help to prevent mistaken conclusions. In his final publication on taxometrics, Meehl (2004) described it this way:

---

John Ruscio, Department of Psychology, The College of New Jersey; Glenn D. Walters, Psychology Services, Federal Correctional Institution, Schuylkill, Minersville, PA; David K. Marcus, Department of Psychology, University of Southern Mississippi; Walter Kacetow, Department of Educational Psychology, Graduate Center of the City University of New York.

The assertions and opinions contained herein are the private views of the authors 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

A crucial feature of the coherent cut kinetics method of taxometrics is reliance on consistency tests to provide multiple lines of evidence about the latent structure. 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 taxonic, one sort of coherent picture will emerge; if it is nontaxonic, a different sort of picture will emerge; if what emerges is unclear, judgment should be suspended until more evidence is examined. (p. 42)

Meehl's approach has three possible outcomes: consistent evidence in favor of a categorical model (which he referred to as *taxonic*), consistent evidence in favor of a dimensional model (which he referred to as *nontaxonic*), or ambiguous evidence from which no conclusion should be drawn.

Perhaps in part because of the availability and intuitive appeal of consistency tests, the taxometric method has become a popular data-analytic tool in psychological assessment research. Table 1 summarizes the 19 studies that have been published in three leading journals in this area (*Assessment*, *Journal of Personality Assessment*, and *Psychological Assessment*), including the constructs, measures, and samples in each investigation. In some studies multiple measures were used to examine the structure of a latent construct (e.g., depression, psychopathy, symptom overreporting), whereas in other studies the focal point was the structure of the construct underlying responses to a particular instrument (e.g., Beck Depression Inventory, Psychopathy Checklist—Revised, Health Problem Overstatement Scale of the Psychological Screening Inventory). This research is being performed and published at an accelerating pace, demonstrating the importance of teasing apart categories and dimensions in psychological assessment.

In addition to documenting this body of research, Table 1 indicates which taxometric procedures and other consistency tests were used in each study. Though the merits of consistency testing have been discussed many times, this practice has not been operationalized. What specific tests should be performed, how should results be combined across multiple tests, and what thresholds should be applied? The taxometric literature is replete with proposed consistency tests, including the many tests introduced by Meehl in his earliest technical reports (e.g., Meehl, 1965, 1968), yet empirically supported guidelines for researchers are few and far between. The reviews of taxometric studies in Haslam and Kim (2002) and Ruscio et al. (2006) reveal that there is little consistency in how researchers have engaged in consistency testing. Table 1 suggests that in the domain of psychological assessment, at least in recent years, researchers have begun to converge on an approach to consistency testing that emphasizes the application of multiple taxometric procedures along with parallel analyses of comparison data as an interpretive aid. Because investigators performing taxometric studies are expected to check the consistency of their findings, this practice needs to be operationalized. There are many possible ways to do so. As a general framework, we suggest that four issues should be addressed empirically to evaluate the merits of any approach to consistency testing. Ideally, the empirical trials should span a wide range of conditions typical of realistic research data.

First, for a candidate index to become a component of the consistency testing process it should differentiate categorical and

dimensional data. An index might involve the output of a single taxometric procedure or of multiple procedures, and it might be calculated objectively or rated subjectively (with appropriate precautions taken to prevent rater bias and to achieve rater agreement). Regardless of the type of index proposed, if it cannot be shown to distinguish data known to be categorical or dimensional then it would be of limited utility in a set of consistency tests.

Second, each candidate index should provide incremental validity. An index that differentiates categorical and dimensional data may be of little use if it does not improve the overall accuracy of decisions relative to what could be achieved without that index. For example, many taxometric estimates of structural model parameters are based in whole or in part on the graphical output of a taxometric procedure. If an index that quantifies the shape of a taxometric graph differentiates categorical and dimensional data well, it is possible that indices based on parameter estimates calculated from the graphical output may provide little or no incremental validity. The inclusion of multiple consistency tests is only desirable to the extent that each makes a nonredundant contribution.

Third, an objective decision rule is required to translate consistency test results into structural conclusions. Even if each test provides an index that differs statistically significantly across categorical and dimensional data sets, some or all of these indices might be of limited practical utility if an objective decision rule is not available, or if its accuracy is poor. There are many ways that the results of multiple tests might be combined to yield a decision. For any index, a single threshold might be applied such that scores at or above the cut are interpreted as evidence of categorical structure and scores below the cut are interpreted as evidence of dimensional structure (or vice versa). Alternatively, dual thresholds might be applied such that an intermediate range of values is set aside as interpretationally ambiguous, with values above or below this range interpreted as for a single threshold. If the results of multiple consistency tests are scaled using the same metric, they might be averaged before applying a single threshold or dual thresholds. Alternatively, one might apply thresholds to each index and require that all decisions are in agreement (a unanimity rule) or that a majority are in agreement (a majority rule).

Fourth, the rates of accurate, ambiguous, and erroneous results should be examined. All else being equal, a decision rule that sets aside more results as interpretationally ambiguous would be expected to achieve a greater accuracy rate among the remaining results. The goal of consistency testing is to avoid both ambiguous and erroneous results, and the balance that each operationalization achieves between these different kinds of undesirable outcomes should be studied.

With this framework in mind, we designed a study to evaluate and refine an operationalization of consistency testing based on fit indices calculated objectively from the graphical output of taxometric procedures. For each taxometric procedure, the fit index is calculated through the parallel analysis of categorical and dimensional comparison data using a four-step process. First, perform a taxometric analysis of the empirical data and retain the graph. If the taxometric procedure provides a panel of curves, average them to provide a single graph. Second, generate populations of comparison data using the two structural models that the taxometric method tests: categorical and dimensional. Based on an analysis of the empirical data, correlations among variables are reproduced

Table 1  
 Summary of 19 Taxometric Studies Published in the Journals *Assessment*, *Journal of Personality Assessment*, and *Psychological Assessment*

Study	Construct	Measure	Sample	Taxometric procedure	CD/CCFI used?	Other consistency test
Strong et al. (1999)	Impression management and self-deception	MMPI-2 Underreporting scales	Child custody litigants	MAMBAC, MAXCOV	No/No	Base rate consistency
Strong et al. (2000)	Symptom overreporting	MMPI-2 Infrequency scales	Psychiatric inpatients and Veterans Affairs Medical Center psychological evaluation patients	MAMBAC, MAXCOV	No/No	Base rate consistency
Franklin et al. (2002)	Depression	MMPI-2 Depression scales	Psychiatric inpatients and outpatients	MAMBAC, MAXCOV	No/No	None
A. M. Ruscio & Ruscio (2002)	Analogue depression	Beck Depression Inventory	Undergraduates	MAMBAC, MAXEIG, L-Mode	Yes/No	Base rate consistency <sup>b</sup> , inchworm consistency test
Strong et al. (2002)	Impression management and self-deception	MMPI-2 Underreporting scales, Balanced Inventory of Desirable Responding—Version 6	Undergraduates, personnel evaluation	MAMBAC, MAXCOV	No/No	Base rate consistency, GFI, Bayes probabilities
Kotov et al. (2005)	Anxiety	Beck Anxiety Inventory, Vulnerability Scale, Anxiety Impairment Scale	Military recruits	MAXCOV, MAXEIG	Yes/No	Base rate consistency, GFI, Bayes probabilities
Strong et al. (2006)	Symptom overreporting	MMPI-2 Infrequency scales	Pretrial forensic examinees adjudicated incompetent to stand trial	MAMBAC, MAXEIG	Yes/No	Base rate consistency <sup>b</sup> , inchworm consistency test
Bernstein et al. (2007)	Anxiety sensitivity	Anxiety Sensitivity Index	Undergraduates and community residents	MAXCOV, MAXEIG	Yes/No	Nose count
Olatunji & Broman-Fulks (2007)	Disgust sensitivity	Disgust Sensitivity Scale, Disgust Emotion Scale	Community, undergraduate, psychiatric outpatients	MAMBAC, MAXEIG, L-Mode	Yes/Yes	Base rate consistency
Walters, Diamond, et al. (2007)	Antisocial personality disorder	Antisocial Features scale of Personality Assessment Inventory	Male and female federal prison inmates	MAMBAC, MAXEIG, L-Mode	Yes/Yes	None
Walters, Duncan, et al. (2007)	Psychopathy	Psychopathy Checklist—Revised	Male federal prison inmates	MAMBAC, MAXEIG, L-Mode	Yes/Yes	None
Walters, Gray, et al. (2007)	Psychopathy	Psychopathy Checklist: Screening Version	Male and female patients and inmates	MAMBAC, MAXEIG, L-Mode	Yes/Yes	None
Broman-Fulks, Green, et al. (2008)	Anxiety sensitivity	Anxiety Sensitivity Profile, Anxiety Sensitivity Index—Revised	Undergraduates	MAMBAC, MAXCOV, MAXEIG, L-Mode	Yes/Yes	Base rate consistency <sup>a</sup>
Broman-Fulks, Hill, & Green (2008)	Perfectionism	Multidimensional Perfectionism Scale, Perfectionism Inventory	Undergraduates	MAMBAC, MAXCOV, MAXEIG, L-Mode	Yes/Yes	Base rate consistency
Parker et al. (2008)	Alexithymia	Toronto Alexithymia Scale	Community, undergraduate, psychiatric outpatients	MAMBAC, MAXCOV, L-Mode	Yes/Yes	None

(table continues)

Table 1 (continued)

Study	Construct	Measure	Sample	Taxometric procedure	CD/CCFI used?	Other consistency test
Walters, Brinkley, et al. (2008)	Psychopathy	Levenson Self-Report Psychopathy Scale	Male and female federal prison inmates	MAMBAC, MAXSLOPE,	Yes/Yes	None
Walters, Rogers, et al. (2008)	Malingering	Structured Interview of Reported Symptoms, MMPI-2	Criminal and civil examinees	MAXEIG, L-Mode MAMBAC, MAXCOV	Yes/Yes	Base rate consistency, case removal, nose count, GFI, Bayes probabilities
J. Ruscio, Brown, & Ruscio (2009)	Major depression	Anxiety Disorders Interview Schedule, Beck Depression Inventory, Depression Anxiety Stress Scales	Undergraduates	MAMBAC, MAXCOV, L-Mode <sup>c</sup>	Yes/Yes	Base rate consistency <sup>b</sup>
Walters et al. (2009)	Exaggerated health complaints	Health Problem Overstatement Scale of Psychological Screening Inventory	Outpatients and forensic clients	MAMBAC, MAXCOV, L-Mode	Yes/Yes	None

*Note.* Studies are presented in chronological order, revealing trends in the methods over time. All authors interpreted the shapes of taxometric curves, and many studies included judgments made by independent raters. In this overview of consistency tests, only data-analytic checks on consistency were recorded. CD = comparison data of some kind were analyzed to facilitate interpretation; CCFI = comparison curve fit index calculated; MIMPI-2 = Minnesota Multiphasic Personality Inventory—Revised; MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; MAXEIG = maximum eigenvalue; L-Mode = latent mode; GFI = goodness-of-fit index; MAXSLOPE = maximum slope.

<sup>a</sup> This test was accompanied by comparison data to facilitate interpretation. <sup>b</sup> The authors noted the limited empirical support for this test. <sup>c</sup> The authors noted that L-Mode was performed, but the results were ambiguous and not reported.

through loadings onto one or more latent factors for the dimensional model. The categorical model consists of two latent classes within which variables may be correlated due to shared loadings onto one or more latent dimensions.<sup>1</sup> Ruscio and Kacetow (2008) provided a program that can be used to generate populations of data that reproduce the observed variables' distributions and correlations using either the categorical or the dimensional model; for categorical data, between-group differences are reproduced as well.

Third, draw multiple random samples from both populations of comparison data, and submit each to the same analysis performed for the empirical data. For each sample of comparison data, retain the graphical results (averaged if necessary, as in Step 1). Fourth and finally, average the results across all samples of categorical comparison data, and then across all samples of dimensional comparison data, to determine which better reproduces the graph for the empirical data. The comparison curve fit index (CCFI) was introduced by Ruscio, Ruscio, and Meron (2007) for this purpose. The fit between results for empirical data and categorical comparison data is quantified using the root-mean-square residual across data points on the curves. The fit between results for empirical data and dimensional comparison data is quantified in the same way. The CCFI is calculated as the ratio of the fit for dimensional comparison data divided by the sum of both fit values. Thus, the CCFI ranges from .00, the strongest support for the dimensional model, to 1.00, the strongest support for the categorical model; a value of .50 is maximally ambiguous in that neither model receives stronger support than the other.

A number of studies have established that the CCFI distinguishes categorical and dimensional data more validly than other indices or approaches to interpreting taxometric results that have been evaluated empirically (Ruscio, 2007; Ruscio & Marcus, 2007; Ruscio et al., 2007; Ruscio & Walters, 2009; Walters & Ruscio, 2009). These studies provide a promising foundation on which to build. With the understanding that other indices or approaches might prove themselves as or more useful in future research, we restricted our attention to the CCFI for four primary reasons. First, the CCFI has performed well in several rigorous studies, effectively differentiating categorical and dimensional data across a wide range of conditions. Second, the CCFI can be calculated for each of several taxometric procedures, which affords the opportunity to examine the consistency of results using a common metric. Third, the CCFI is based exclusively on the graphical output that forms the centerpiece of taxometric analysis. Therefore, the CCFI conveniently supplements the presentation and inspection of taxometric curves with an objective index that can serve as an interpretive aid. Fourth, 12 of the 19 studies in Table 1 used the CCFI, including 12 of the 13 studies published in these assessment journals since 2007.

<sup>1</sup> In the taxometric literature, the latter model is usually referred to as *taxonic*, and the two latent classes as the *taxon* and its *complement*. The purely dimensional model often is referred to as *nontaxonic*, a term whose ambiguity is resolved by the context within which it is used. For example, Waller and Meehl (1998) compared the taxometric results for taxonic data to those for "nontaxonic" data in several chapters, after having established that the latter refers to a dimensional structural model.

The present study was designed to examine six related issues in the development, refinement, and evaluation of operationalizations of consistency testing. First, for the same target data sets, how well does the CCFI for each of several taxometric procedures perform? We compare accuracy rates for a large number of samples that span a wide range of data conditions. Second, what are the consequences of using various thresholds for interpreting the CCFI? We examine results obtained using a single threshold at CCFI = .50 as well as using dual thresholds that diverge symmetrically around this value (thresholds at CCFIs of .45 and .55 and of .40 and .60). Third, does using multiple populations of categorical comparison data improve the performance of the CCFI? We evaluate the potential benefits of generating categorical comparison data using each of several estimates of the two groups' relative sizes. Fourth, how should agreement across results from multiple taxometric procedures be defined? We compare the performance of a mean CCFI value as well as a unanimity rule and a majority rule for deciding between categorical, dimensional, and ambiguous outcomes. This comparison also explores the trade-off between the frequencies of ambiguous and erroneous results. Fifth, which taxometric procedures provide incremental validity? We examine the redundancy of CCFI values across procedures as well as the extent to which each provides incremental validity when included in a decision rule along with others. Sixth, how accurate can researchers expect results to be? We present the probability that the CCFI—for a single procedure or the mean of several procedures—is correct across its full range of values as well as its robustness across data conditions.

## Method

### Design and Data Generation

A total of 100,000 categorical and dimensional data sets (50,000 for each structure) were generated using a Monte Carlo design in which data parameters were independently randomly sampled from specified ranges. These data conditions were identical to those of Ruscio and Kacetow (2009) and Ruscio and Walters (2009), which in turn were an expanded version of those of Ruscio (2007, Study 2) and Ruscio et al. (2007, Study 3). For categorical data, which consisted of two latent classes referred to as the *taxon* (higher scoring group) and its *complement* (lower scoring group), random values were drawn for the following parameters of each target data set: sample size ( $N = 300$  to 1,000), number of indicators ( $k = 3$  to 8), taxon base rate ( $p = .10$  to .50), indicator validity (standardized mean difference between classes of  $d = 1.25$  to 2.00), within-group correlation ( $r = .00$  to .30), asymmetry ( $g = .00$  to .30), tail weight ( $h = .00$  to .15), and variance ratio ( $VR = .25$  to 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 that  $VR = X$ , and with probability .50 that  $VR = 1/X$ . The values of  $g$  and  $h$  were used to generate data from a  $g$ -and- $h$  distribution (Hoaglin, 1985, p. 486), with a random normal variable  $Z$  transformed as follows:

$$y_{g,h}(Z) = \left( \frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2}$$

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, the smallest skew ( $\gamma_1$ ) and kurtosis ( $\gamma_2$ ) 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  $\gamma_1$  and  $\gamma_2$  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 Ruscio and Kacetow (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 were randomly selected and identified as taxon members, with the remainder identified as members of the complement class. The variance ratio was achieved by multiplying scores in the taxon by  $X$  (when  $VR > 1$ ) or  $1/X$  (when  $VR < 1$ ). 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 using 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 Ruscio and Kacetow (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.

When generating each of the 100,000 samples of target data for the present study, the random number seed was initialized to the sample number. This allows the exact reproduction of these samples by anyone who would like to replicate the results or perform new analyses of the same data. An R file containing programs to reproduce our data is available on request.<sup>2</sup>

<sup>2</sup> The 100,000 data files themselves would be too cumbersome to upload to a website or to transmit electronically as an e-mail attachment. To estimate the total file size, we wrote the first 100 samples to individual, tab-delimited text files. This required nearly 6MB of disk space, suggesting that the full collection of samples would approach 6GB. Saving to a compressed (.zip) folder cut this size nearly in half, but this remains unwieldy. We recommend that interested parties run the comparatively small R programs we can provide to recreate some or all of these 100,000 data files on demand.

## Data Analysis

**Taxometric procedures.** For each sample, we performed the four procedures used most commonly in published taxometric studies (Ruscio et al., 2006, p. 280): MAMBAC, MAXCOV, MAXEIG, and L-Mode. Ruscio et al. (2006) discussed these and other taxometric procedures, along with the choices one must make to implement them. Below, we briefly describe each procedure, cite a primary source, and explain how it was implemented in the present study.

The MAMBAC (mean above minus below a cut; Meehl & Yonce, 1994) procedure assigns one variable to the role of “input” indicator and another to the role of “output” indicator. Cases are sorted according to their scores on the input indicator, which is then used to locate a series of cutting scores. We located 50 evenly spaced cutting scores beginning and ending 25 cases from each extreme. The mean difference on the output indicator is calculated for scores falling above and below each cutting score. A MAMBAC graph plots the series of mean differences, on the ordinate, by the location of each cutting score along the series of cases as sorted by the input indicator. For categorical data, the prototypical MAMBAC graph exhibits a peak near the cutting score that optimally separates the two groups. For dimensional data, the prototypical MAMBAC graph is concave rather than peaked. With  $k \geq 2$  variables, one can perform a total of  $k(k - 1)$  MAMBAC analyses by assigning variables to all possible indicator configurations. We averaged the results for all MAMBAC analyses of a sample to yield a single graph.

The MAXCOV (maximum covariance; Meehl & Yonce, 1996) procedure assigns one variable to the role of input indicator and two variables to the role of output indicators. Cases are sorted according to their scores on the input indicator, which is then used to form a series of ordered subsamples. We constructed a series of 50 overlapping windows such that all windows had the same  $n$  and adjacent windows overlapped 90% with one another. The covariance of two output indicators is calculated within each subsample. A MAXCOV graph plots the series of covariances, on the ordinate, by the mean score for cases within each ordered subsample along the input indicator. For categorical data, the prototypical MAXCOV graph exhibits a peak near the score that optimally separates the two groups. For dimensional data, the prototypical MAXCOV graph is flat rather than peaked. With  $k \geq 3$  variables, one can perform a total of  $k(k - 1)(k - 2)/2$  MAXCOV analyses by assigning variables to all possible indicator configurations. We averaged the results for all MAXCOV analyses of a sample to yield a single graph.

The MAXEIG (maximum eigenvalue; Waller & Meehl, 1998) procedure assigns one variable to the role of input indicator and all remaining variables to the role of output indicators. Overlapping windows can be formed in the same way as for MAXCOV, which is what we chose to do. Rather than calculating the covariance of two output indicators within each subsample, MAXEIG involves the calculation of the first (largest) eigenvalue of the covariance matrix for all output indicators. The covariance matrix is a modified variance-covariance matrix with 0s placed on the diagonal so that variances are removed and only covariances remain. A MAXEIG graph is plotted and interpreted just like a MAXCOV graph. With  $k \geq 3$  variables, one can perform a total of  $k$  MAXEIG analyses by assigning each variable to the role of input indicator once. We averaged the results for all MAXEIG analyses of a sample to yield a single graph.

The L-Mode (latent mode; Waller & Meehl, 1998) procedure begins with a principal axis factor analysis of all available variables. Scores on a single factor are calculated using Bartlett’s (1937) method. An L-Mode graph is the factor score density plot. For categorical data, the prototypical L-Mode graph is bimodal, with each mode corresponding to the typical factor score for one group. For dimensional data, the prototypical L-Mode graph is unimodal. With  $k \geq 3$  variables, one can perform a single L-Mode analysis that includes all variables, and that is what we chose to do.

**Comparison data.** For each sample, three populations of categorical comparison data were generated, along with one population of dimensional comparison data. Creating categorical comparison data requires the assignment of cases to groups, which can be done in multiple ways. To examine the potential benefit of using multiple populations of categorical comparison data, three such populations were generated for each sample. Because cases are not assigned to groups to create dimensional comparison data, only a single population of this type was formed.

To generate each population of categorical comparison data, the base-rate classification technique was used to assign cases to groups. Prior research suggests that this approach yields good classification accuracy (Ruscio, 2009) and that CCFI values calculated using this type of comparison data differentiate categorical and dimensional target data well (see Ruscio, 2007; Ruscio & Marcus, 2007; Ruscio et al., 2007; Ruscio & Walters, 2009; Walters & Ruscio, 2009). The mean of the taxon base rates estimated from the MAXCOV and MAXEIG analyses was used, along with two additional base rates. To cover a modest range of base rates close to the value estimated using the taxometric results, the series of three base rates differed from one another by .05. For example, if the mean of the MAXCOV and MAXEIG estimates was .17, populations of categorical comparison data were generated using base rates of .12, .17, and .22. The lowest base rate in the series was not allowed to drop below .05 (e.g., if the mean of the MAXCOV and MAXEIG estimates was .08, then populations of categorical comparison data were generated using base rates of .08, .13, and .18). Each base rate was used to classify cases into 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.

**Comparison curve fit index (CCFI).** Once all four populations of comparison data (three categorical and one dimensional) were generated, taxometric analyses were performed for each of  $B = 10$  samples drawn randomly from each population. Analyses were implemented in the same way as for the target data. Using the results for the sample of target data and its corresponding samples of comparison data, the CCFI was calculated as described earlier. For MAMBAC, MAXCOV, and MAXEIG, fit was calculated across the 50 data points on each curve. For L-Mode, the calculation was adapted to maximize the fit, assessed by Euclidean distance, for each type of comparison data by allowing a horizontal shift to better align the curves. Details are provided by Ruscio and Walters (2009), who introduced this modification and found that it improved the CCFI considerably for this procedure.

**Results**

**Multiple Populations of Categorical Comparison Data**

The first question we examined with the CCFI values provided by each taxometric procedure for the 100,000 samples was whether creating three populations of categorical comparison data increased accuracy beyond what could be achieved using a single population. To address this question, we calculated the percentage of all samples classified correctly by each taxometric procedure. A single threshold at .50 was applied, with values at or above this level indicative of categorical data and values below this level indicative of dimensional data. This percentage was calculated first using CCFI values based on a single population of categorical comparison data—generated using the taxon base rate estimated from the analysis—and then using the mean of the three CCFI values based on all three populations of categorical comparison data. Though there was variation in accuracy rates across procedures, there was never more than a 0.5% difference in accuracy rates for the same procedure with three versus one population of categorical comparison data. Because accuracy was always higher when three populations were used, only these mean CCFIs are examined in subsequent analyses.

**Redundancy Among Taxometric Procedures**

The next question we examined was whether each procedure provided independent evidence, or whether any procedures' results

were highly redundant. This was tested by correlating the four procedures' CCFI values across all samples. Though all correlations were large ( $r > .85$ ), only that between MAXCOV and MAXEIG suggested nearly complete redundancy:  $r > .999$ . Not only were these CCFIs almost perfectly correlated, but also there was no detectable mean difference ( $M_D = .000, SD_D = .006$ ). Not surprisingly, the overall accuracy of these indices was similar, too: 92.34% for MAXCOV and 92.28% for MAXEIG. Because of their redundancy, we decided to retain only one of these procedures' results for subsequent analysis. We selected MAXCOV because of its higher accuracy, but we emphasize that the similarities in these procedures' results were much more striking than the differences.

**Accuracy of Procedures Individually or in Combination**

**Individual procedures.** The first three blocks in Table 2 display the accuracy rates for the MAMBAC, MAXCOV, and L-Mode procedures. This is listed for all 100,000 samples as well as separately for the categorical and the dimensional samples. Further, accuracy rates are shown for a single CCFI threshold at .50, dual thresholds of .45 and .55, and dual thresholds of .40 and .60. For dual thresholds, CCFI values below the lower threshold were indicative of dimensional data, values above the upper threshold were indicative of categorical data, and values at or between the thresholds were treated as ambiguous and no structural conclusion was reached. Table 2 presents the accuracy rates,

Table 2  
*Accuracy Rates (and Percentage of Interpretable Results) for CCFIs of All Three Procedures*

Procedure	Threshold	All samples	Categorical samples	Dimensional samples
MAMBAC	.50	96.3 (100.0)	99.4 (100.0)	93.3 (100.0)
	.45, .55	98.2 (94.5)	99.8 (98.9)	96.5 (90.0)
	.40, .60	99.2 (87.6)	99.9 (97.3)	98.3 (77.9)
MAXCOV	.50	92.3 (100.0)	90.1 (100.0)	94.6 (100.0)
	.45, .55	95.8 (89.5)	93.9 (89.1)	97.8 (89.9)
	.40, .60	97.9 (77.1)	96.7 (78.1)	99.1 (75.8)
L-Mode	.50	94.7 (100.0)	93.1 (100.0)	96.3 (100.0)
	.45, .55	97.3 (92.0)	96.3 (91.2)	98.3 (92.8)
	.40, .60	98.9 (81.3)	98.4 (82.2)	99.3 (80.4)
Mean	.50	98.0 (100.0)	97.7 (100.0)	98.3 (100.0)
	.45, .55	99.4 (94.8)	99.3 (95.3)	99.6 (94.3)
	.40, .60	99.9 (86.0)	99.8 (89.3)	99.9 (82.8)
Unanimity	.50	99.7 (86.7)	99.6 (86.9)	99.8 (86.6)
	.45, .55	99.9 (76.0)	99.9 (78.7)	99.9 (73.4)
	.40, .60	100.0 (61.2)	100.0 (69.0)	100.0 (63.4)
Majority	.50	97.2 (100.0)	96.4 (100.0)	98.0 (100.0)
	.45, .55	98.9 (94.1)	98.4 (94.0)	99.4 (94.2)
	.40, .60	99.7 (85.0)	99.5 (87.2)	99.8 (82.8)

*Note.* With a single threshold of .50, data were identified as dimensional if  $CCFI < .50$  and categorical if  $CCFI \geq .50$ . For dual thresholds of .45 and .55, data were identified as dimensional if  $CCFI < .45$ , ambiguous if  $.45 \leq CCFI \leq .55$ , and categorical if  $CCFI > .55$ . For dual thresholds of .40 and .60, data were identified as dimensional if  $CCFI < .40$ , ambiguous if  $.40 \leq CCFI \leq .60$ , and categorical if  $CCFI > .60$ . The percentage of all samples with interpretable (i.e., not ambiguous) results is indicated in parentheses after each accuracy rate. CCFI = comparison curve fit index; MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; L-Mode = latent mode; Mean = mean of all three CCFI values; Unanimity = all three CCFIs met specified criteria (i.e., CCFIs outside an intermediate range of values set aside as ambiguous); Majority = CCFIs for any two procedures met specified criteria.

with the percentage of samples yielding interpretable (nonambiguous) results in parentheses.

Across all samples, MAMBAC was most accurate (96.3% correct with a single threshold), MAXCOV was least accurate (92.3%), and L-Mode was intermediate (94.7%). This pattern held for the categorical samples. For the dimensional samples, L-Mode was most accurate, followed by MAXCOV and then MAMBAC. To put it another way, L-Mode was more accurate than MAXCOV for categorical and dimensional data, but MAMBAC was more accurate than both for categorical data and less accurate than both for dimensional data. All of these trends remained the same when accuracy was assessed using dual thresholds. To compare accuracy across procedures independent of threshold, the probability-based measure of effect size  $A$  was used (Ruscio, 2008). This is equivalent to the area under a receiver operating characteristic curve, and in this context it can be interpreted as the probability that a randomly chosen categorical sample yielded a larger CCFI than a randomly selected dimensional sample. For MAMBAC, MAXCOV, and L-Mode,  $A = .998$ ,  $.976$ , and  $.988$ , respectively. This suggests that the CCFI differentiates categorical and dimensional data exceptionally well for each of these procedures.

**Multiple procedures.** We examined three different ways of combining MAMBAC, MAXCOV, and L-Mode results. First, one can calculate their mean CCFI value and then apply either a single threshold or dual thresholds. Second, one can apply single or dual thresholds to each CCFI, and then require that all three results agree with one another. For example, using dual thresholds, all three CCFIs need to fall below the lower threshold to indicate dimensional data, all three CCFIs need to fall above the upper threshold to indicate categorical data, and in any other instance the results would be considered ambiguous. Whereas this second decision rule requires unanimity, a third rule relaxes this by requiring that a majority of CCFI values (two of three) fall below the lower threshold or above the upper threshold to reach a structural conclusion. The bottom three blocks of Table 2 present the accuracy rates for these three ways to combine the results of multiple taxometric procedures. Accuracy is shown for all samples, for categorical data, and for dimensional data, with the percentage of samples yielding interpretable results in parentheses. Figure 1 displays the results for all samples graphically, with the best configuration of results yielding data points near the upper-left corner of each plot (high accuracy with few ambiguous results).

These analyses reveal several findings. First, each of the three methods of combining results across procedures achieved greater accuracy than any individual procedure—holding constant the percentage of interpretable results. In other words, if one would like to reach conclusions for a certain number of samples, neither MAMBAC, MAXCOV, nor L-Mode attained as high an accuracy level as any of the decision rules that included all three procedures' CCFIs. Second, the use of a single mean CCFI value outperformed the majority and unanimity rules. Holding constant the level of interpretable results, the mean CCFI achieved the highest accuracy level of all; expressed as a measure of effect size,  $A = .998$  for the mean CCFI. Third, requiring unanimity across the three CCFIs can provide an incredibly stringent operationalization of consistency testing. For example, when dual thresholds of  $.40$  and  $.60$  were applied, the unanimity rule was correct for 61,186 of the 61,203 samples with interpretable results, or an accuracy rate of 99.97%.

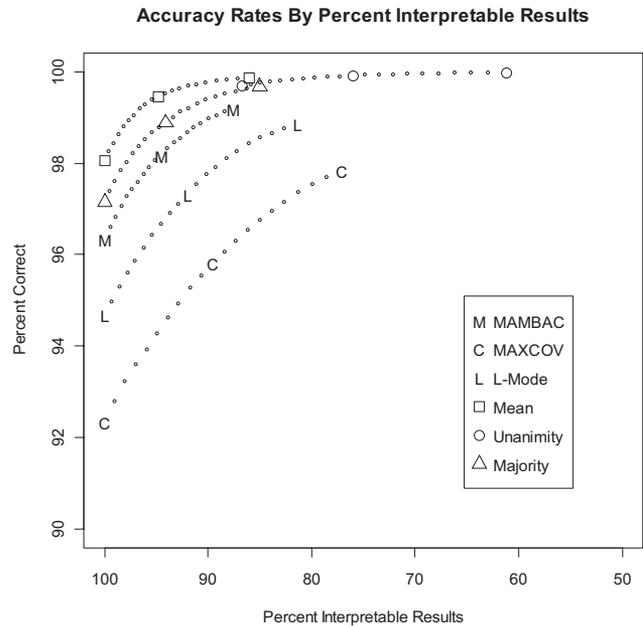


Figure 1. The graph shows the trade-off between accuracy and the percentage of samples meeting criteria. Symbols shown in the legend are plotted for the single threshold of  $.50$  (data point on the left end of each curve), dual thresholds of  $.45$  and  $.55$  (data point in the middle of each curve), and dual thresholds of  $.40$  and  $.60$  (data point on the right end of each curve). Data points fleshing out each curve show results for a series of thresholds that varied in increments of  $.005$ . CCFI = comparison curve fit index; MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; L-Mode = latent mode; Mean = mean of CCFIs for all three procedures; Unanimity = CCFIs for all three procedures met specified criteria (i.e., CCFIs outside an intermediate range of values set aside as interpretationally ambiguous); Majority = CCFIs for any two procedures met specified criteria.

**Incremental validity.** The fact that accuracy was greater using the results of all three procedures than using any individual procedure means that incremental validity is worth examining closely. Table 3 shows the results of separate series of analyses in which the results of each procedure were combined with those of others in all possible combinations and orders. These results are shown for several thresholds as well as two ways of combining CCFIs from more than one procedure (mean CCFI, unanimity rule). Figure 2 plots the findings for the mean CCFIs, which makes it easier to see at a glance when incremental validity was observed. Examining all of this information reveals a clear trend: MAMBAC and L-Mode always improved performance when their CCFI was included, but the same was not true of MAXCOV. When the CCFIs for all three procedures were combined, the results were no better than when the CCFIs for only MAMBAC and L-Mode were combined.

### Probability of Correctly Differentiating Categorical and Dimensional Data

To this point, accuracy has been compared across procedures, thresholds, and methods of combining multiple CCFIs. Another

Table 3  
Accuracy Rates (and Percentage of Interpretable Results) for CCFIs of One, Two, or Three Procedures

Procedure	Method of combination	Single threshold: .50	Dual thresholds: .45 and .55	Dual thresholds: .40 and .60
MAMBAC		96.3 (100.0)	98.2 (94.5)	99.2 (87.6)
MAXCOV		92.3 (100.0)	95.8 (89.5)	97.9 (77.1)
L-Mode		94.7 (100.0)	97.3 (92.0)	98.9 (81.3)
MAMBAC & MAXCOV	Mean	97.6 (100.0)	99.2 (94.3)	99.8 (85.8)
MAMBAC & L-Mode	Mean	98.1 (100.0)	99.4 (95.2)	99.8 (87.4)
MAXCOV & L-Mode	Mean	95.6 (100.0)	98.1 (92.1)	99.3 (80.5)
MAMBAC, MAXCOV, & L-Mode	Mean	98.0 (100.0)	99.4 (94.8)	99.9 (86.0)
MAMBAC & MAXCOV	Unanimity	99.1 (90.4)	99.7 (81.3)	99.9 (68.7)
MAMBAC & L-Mode	Unanimity	99.5 (92.0)	99.8 (84.0)	99.9 (72.1)
MAXCOV & L-Mode	Unanimity	97.8 (91.1)	99.1 (80.9)	99.7 (66.7)
MAMBAC, MAXCOV, & L-Mode	Unanimity	99.7 (86.7)	99.9 (76.0)	100.0 (61.2)

Note. With a single threshold of .50, data were identified as dimensional if  $CCFI < .50$  and categorical if  $CCFI \geq .50$ . For dual thresholds of .45 and .55, data were identified as dimensional if  $CCFI < .45$ , ambiguous if  $.45 \leq CCFI \leq .55$ , and categorical if  $CCFI > .55$ . For dual thresholds of .40 and .60, data were identified as dimensional if  $CCFI < .40$ , ambiguous if  $.40 \leq CCFI \leq .60$ , and categorical if  $CCFI > .60$ . The percentage of all samples with interpretable (i.e., not ambiguous) results is indicated in parentheses after each accuracy rate. CCFI = comparison curve fit index; MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; L-Mode = latent mode; Mean = mean of all three CCFI values; Unanimity = CCFIs for all two or more procedures met specified criteria (i.e., CCFIs outside an intermediate range of values set aside as ambiguous).

way to examine accuracy is to pose the question of how likely a given CCFI value is to correctly identify categorical versus dimensional data. Figure 3 plots the probability that a CCFI value—specifically, the mean of the MAMBAC, MAXCOV, and L-Mode procedures—is correct. A solid vertical reference line shows the CCFI value of .50, which theoretically should be maximally ambiguous, and dotted vertical reference lines show the dual CCFI thresholds of .40 and .60. Toward the bottom of the graph, the densities of CCFI values for dimensional and categorical data are plotted. Toward the top of the graph, the probability that the CCFI was correct is plotted; this was calculated as the height of the

density for the correct model (dimensional data for  $CCFIs < .50$ , categorical data for  $CCFIs \geq .50$ ) divided by the sum of the heights of both densities.

As expected, the densities intersected near the theoretically optimal threshold of  $CCFI = .50$  and the probability of correct structural model identification was fairly symmetric about this point. Accuracy remained quite high as CCFI values approached .50 from both directions. For mean CCFI values beyond the dual thresholds of .45 and .55, probabilities usually remained above .90, and beyond the dual thresholds of .40 and .60, the probabilities were close to 1.00.

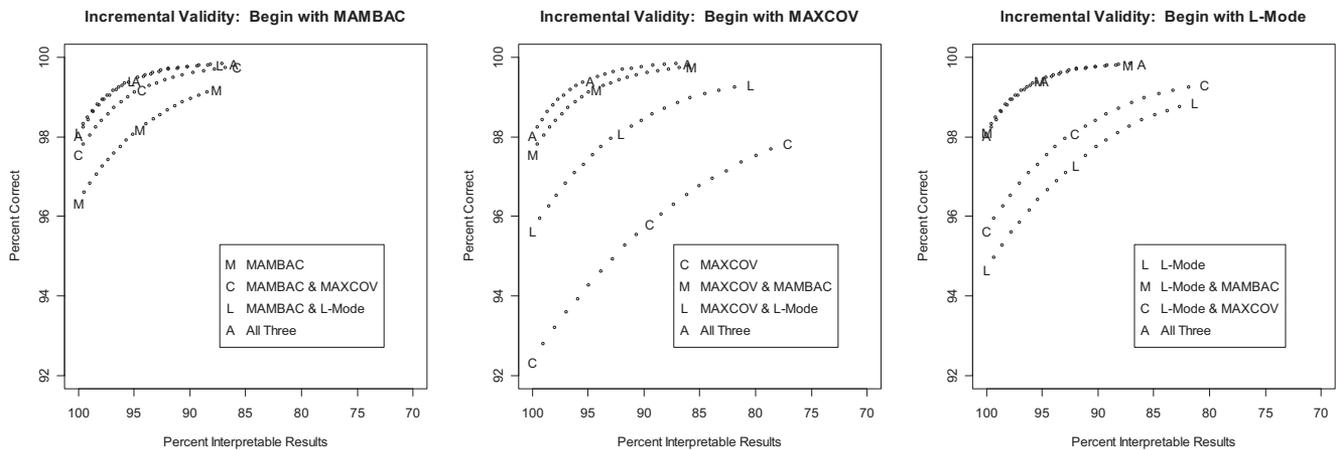


Figure 2. Results for two or three procedures are based on their mean comparison curve fit index (CCFI) values. Each graph shows the trade-off between accuracy and the percentage of samples meeting criteria. Symbols shown in the legend are plotted for the single threshold of .50 (data point on the left end of each curve), dual thresholds of .45 and .55 (data point in the middle of each curve), and dual thresholds of .40 and .60 (data point on the right end of each curve). Data points fleshing out each curve show results for a series of thresholds that varied in increments of .005. MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; L-Mode = latent mode.

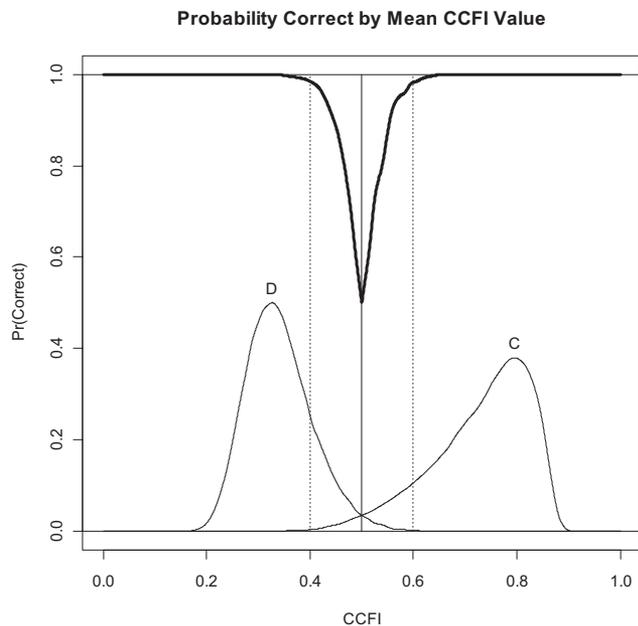


Figure 3. At the bottom of the graph, the densities of comparison curve fit index (CCFI) values for dimensional (D) and categorical (C) samples are plotted as thin lines; their heights were rescaled to fit within the lower half of the graph region. Using these densities, the probability that the CCFI correctly identified categorical or dimensional structure,  $\text{Pr}(\text{Correct})$ , was calculated across the full range of CCFI values; this is plotted as a thick line in the upper region of the graph. The vertical dotted lines represent broad dual thresholds at  $\text{CCFI} = .40$  and  $\text{CCFI} = .60$ .

### Accuracy Levels Across Data Conditions

Across all 100,000 samples, using the mean CCFI with dual thresholds of .45 and .55 achieved an excellent balance between achieving interpretable results (94.8% of all samples) and accurately distinguishing categorical and dimensional data (99.4% correct). How does this performance vary across data conditions? The graphs in Figure 4 plot accuracy as a function of study design factors, which were broken down into eight levels apiece to flesh out the shape of each curve while retaining as many samples as possible within each level; the number of indicators ( $k$ ) varied only across six discrete values of 3 to 8. Triangles pointing upward show the percentage of samples whose structure was identified correctly ( $\text{CCFI} < .45$  for dimensional data or  $\text{CCFI} > .55$  for categorical data), triangles pointing downward show the percentage of errors ( $\text{CCFI} > .55$  for dimensional data or  $\text{CCFI} < .45$  for categorical data), and open circles show ambiguous results ( $.45 \leq \text{CCFI} \leq .55$ ). An additional series of data points, plotted as filled circles, represents the percent correct out of all interpretable results (percent correct plus percent incorrect), or how often the CCFI was correct after ambiguous results were set aside.

Perhaps the most striking trend across all of these graphs is that the filled circles remained close to 100% across all levels of all design factors. Even though the percent correct sometimes dropped well below 100% (e.g., with relatively low indicator validity or relatively large within-group correlations), the offsetting increase was observed among ambiguous results to a much greater extent than erroneous results. Throughout the levels for design factors

plotted in Figure 4, the percent incorrect was always less than 3%, and usually less than 1%; the percent correct among interpretable results usually exceeded 99%.

Figure 4 shows accuracy across the data conditions used to generate the 100,000 samples, with results plotted separately for categorical and dimensional data. Figure 5 extends this analysis by plotting accuracy for all samples, with data conditions that are known to researchers (sample size and number of indicators) or estimated from the data. In other words, whereas researchers working with empirical data will not know the data parameters (e.g., indicator validity in the population), they can estimate them from their data. In the present study, four indices were estimated by assigning cases to groups as was done to generate categorical comparison data. First, the taxon base rate was calculated as the proportion of cases in the higher scoring group. Second, indicator validity was calculated as the mean of each indicator's standardized mean difference (Cohen's  $d$ ) across groups. Third, within-group correlation was calculated as the mean of the indicator intercorrelations within each group. Fourth and finally, an index referred to as the correlational difference was calculated as the mean indicator intercorrelation in the full sample minus the mean within-group correlation. Figure 5 shows the percent correct, incorrect, and ambiguous across levels of each of these data conditions.

The first noteworthy aspect of these graphs is that in many instances, one or more of the estimated data conditions fell outside the bounds considered acceptable for taxometric analysis (e.g., indicator validity  $d < 1.25$ , within-group correlation  $r > .30$ ). Nonetheless, the percent incorrect never exceeded 1.5% and the percent correct among interpretable results usually remained at or above 99%. Whereas there was little relationship between estimated taxon base rate and the likelihood of obtaining ambiguous results, fewer ambiguous results were associated with a larger sample size, number of indicators, estimated indicator validity, or estimated correlational difference. As anticipated by Ruscio et al. (2006), the trend toward smaller estimated differences between full-sample and within-group correlations was more strongly associated with ambiguous results than the trend toward either smaller or larger estimated within-group correlations.

### Testing the Limits

The data conditions in this study conform to the limits for taxometric analysis recommended by Meehl (1995). We performed a final series of tests to examine what happens as data conditions exceed these limits. Specifically, Meehl suggested that taxometric results should be trustworthy provided that  $N \geq 300$ ,  $p \geq .10$ ,  $d \geq 1.25$ , and  $r$  (within group)  $\leq .30$ . These rules of thumb appear to reflect Meehl's 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 using the results of the MAXCOV taxometric procedure included data conditions that reached beyond the limits recommended by Meehl. Performance decrements were observed, suggesting that the parameter estimates provided by this procedure degraded near Meehl's limits. More recently, Ruscio and Walters (2009) observed decrements in the performance of the L-Mode procedure for data conditions outside these limits.

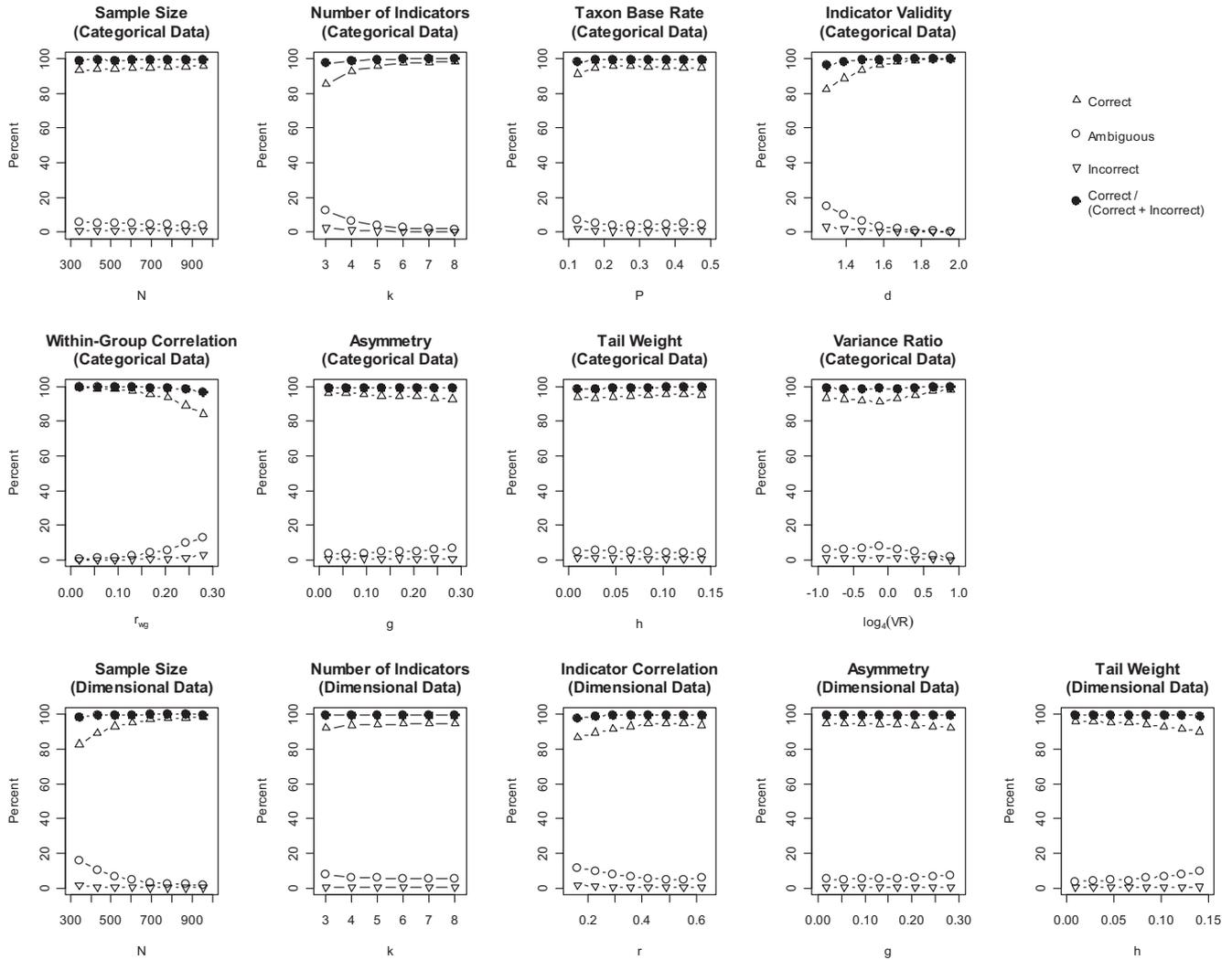


Figure 4. Accuracy rates for mean comparison curve fit index (CCFI) values by levels of each factor in the simulation study design. The percentage of incorrect values remained below 3.0% in all data conditions examined.

Following the lead of Ruscio and Walters (2009), we extended the findings of the present study by generating 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 rules of thumb:  $.05 < p < .10$ ,  $0.80 < d < 1.25$ ,  $.30 < r < .50$ , and  $100 \leq N \leq 300$ . Because Meehl offered no rules of thumb for dimensional data and because none of the relevant factors were associated with a precipitous decline in our results, we examined only  $100 \leq N \leq 300$ . For each of these five new conditions, we generated 2,000 new samples. The results for these 10,000 new samples are presented along with those for our original samples in Figure 6, which extends the range for each factor by replotting the original beside the new results; a dotted line indicates Meehl’s 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.

For categorical data, there was a steep decline in accuracy when  $d < 1.25$  or  $r > .30$ , and the CCFI did not afford good protection in this region. The decline was more modest for data with base rates below  $p = .10$  and very slight for data with  $N < 300$ . For dimensional data, there was a sharper decline for data with  $N < 300$ . These findings underscore the importance of paying careful attention to the adequacy of the data for taxometric analysis. Venturing beyond the recommended limits can significantly increase the risk of obtaining ambiguous or inaccurate results, and operationalizing consistency testing using the CCFI provides limited protection against this risk.

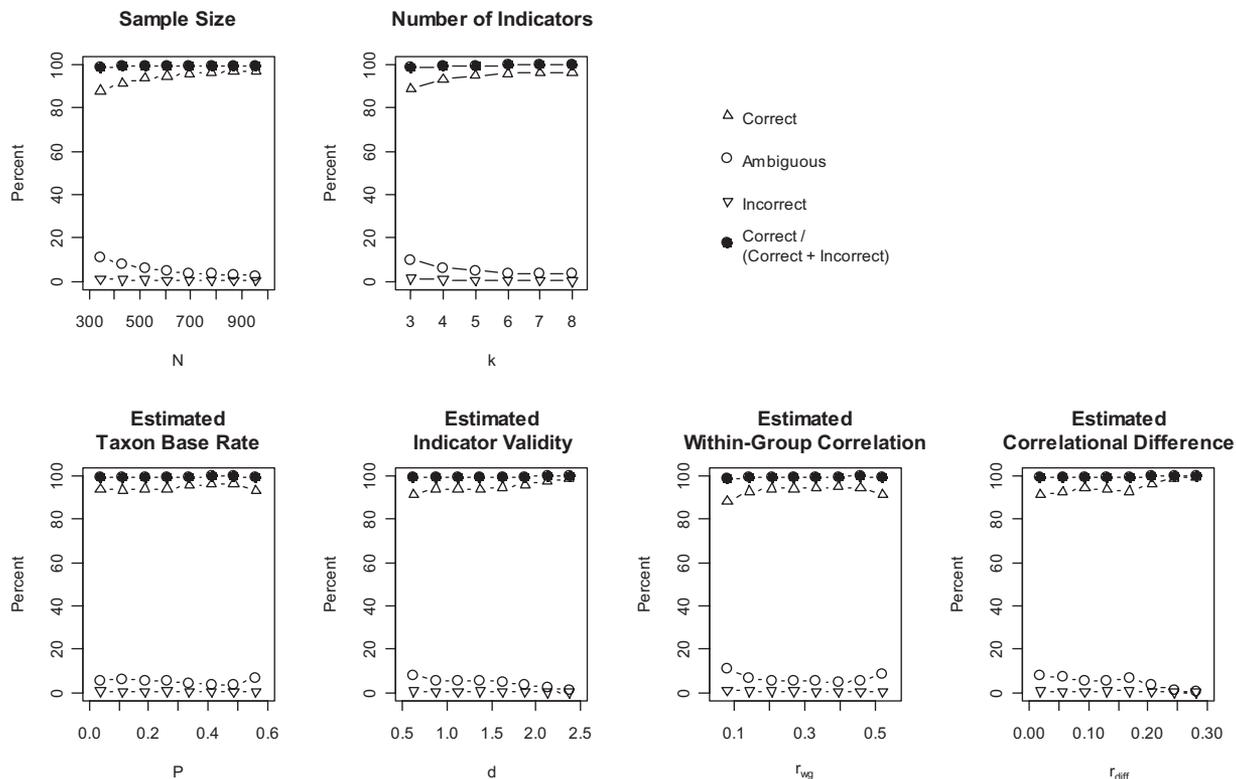


Figure 5. Accuracy rates for mean comparison curve fit index (CCFI) values by sample size, number of indicators, and four estimated data parameters. The taxon base rate was estimated using MAXCOV (maximum covariance) and MAXEIG (maximum eigenvalue) analyses, and the mean of these estimates was used to assign cases to groups with the base-rate classification method. Estimated indicator validity was calculated as the mean value of Cohen’s *d* for all indicators across groups. Estimated within-group correlation was calculated as the mean correlation among all indicators within group. Estimated correlational difference was calculated as the difference between the mean indicator correlation in the full sample minus the mean indicator correlation within group. The percentage of incorrect values remained below 1.5% in all data conditions examined.

**An Empirical Illustration**

As a concrete example of taxometric consistency testing, we examined the latent structure of adolescent externalizing disorder using indicators constructed from self-report measures of antisocial behavior, substance misuse, tobacco use, and sexual experience. Externalizing disorder was originally designed to serve as a general term for disruptive childhood disorders like attention-deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder, but it has recently been extended to adult antisocial behavior and substance abuse (Krueger, Markon, Patrick, & Iacono, 2005). Krueger et al. (2005) studied adult externalizing disorder in adult twin pairs from the Minnesota Twin Family Study using latent class and latent trait analysis and concluded that the evidence supported a dimensional latent structure. We performed taxometric analyses to illustrate the empirically supported use of consistency tests developed in our simulation study. At the outset, we emphasize that these analyses are provided for illustrative purposes only and not to expand upon or test the generality of the results from the simulation study.

A large sample of adolescents from the 2007 Youth Risk Behavior Surveillance (YRBS; Eaton et al., 2008) provided the data for these

analyses. The YRBS is a nationally representative sample of students in Grades 9–12 attending public and private schools in the 50 states and the District of Columbia. Of the 14,041 usable questionnaires collected as part of the 2007 YRBS, 11,203 (79.8%) had complete data on the eight questions used as indicators in these analyses. Among these individuals, 51.8% were female and 48.2% were male, the average age was 16.14 years, and the ethnic breakdown was 44.6% White, 20.0% Black, 26.9% Hispanic, 3.2% Asian, and 5.3% other.

Eight self-report items were selected from the YRBS database to represent potential externalization problems, and these items covered antisocial behavior, alcohol and drug use, and sexual experience. The two antisocial behavior indicators, along with endpoints in their coding schemes,<sup>3</sup> were “carried a weapon (gun, knife, club) in the last 30 days” (1 = *no days* to 5 = *six or more days*) and “involved in a physical fight in the last 12 months” (1 = *no times* to 8 = *12 or more times*). The four alcohol and drug items were “number of days smoked cigarettes in the last 30 days” and

<sup>3</sup> Full coding schemes for all items are available on request.

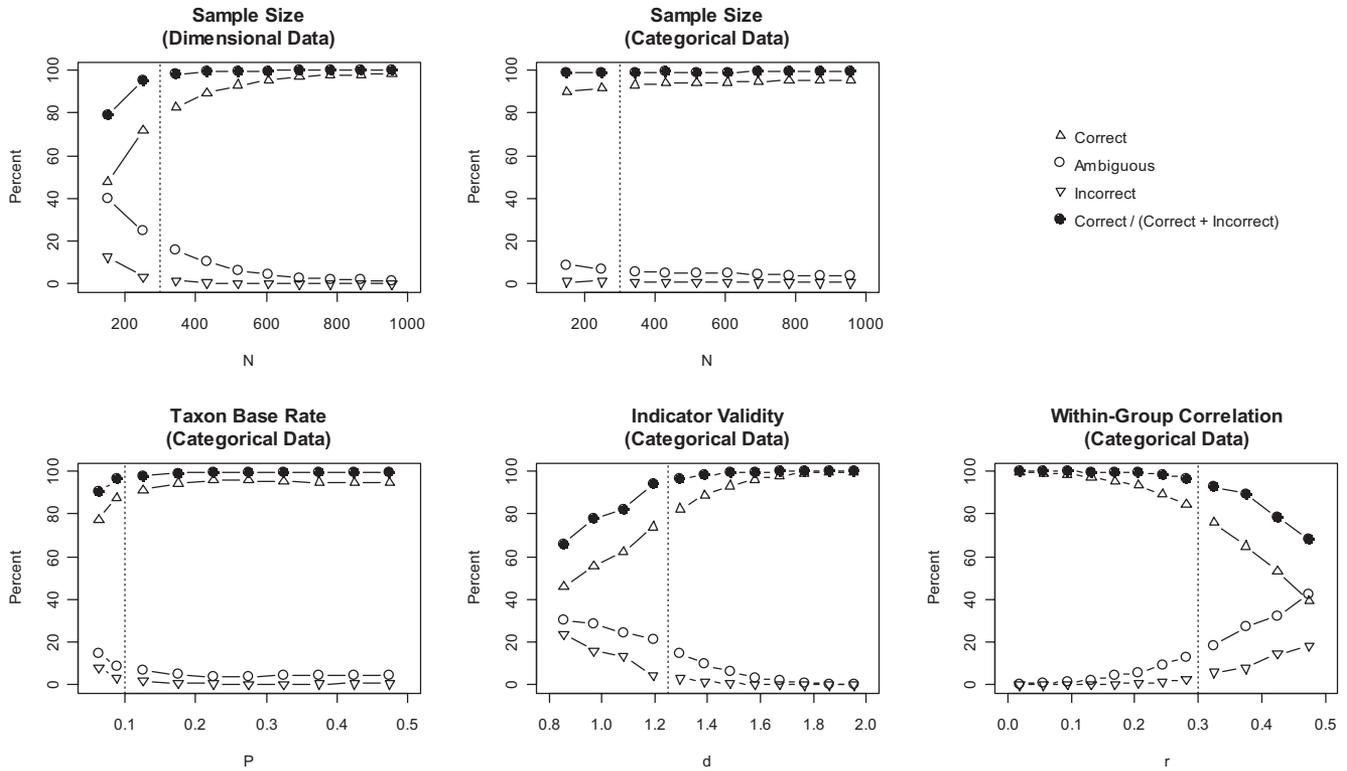


Figure 6. Accuracy rates for mean comparison curve fit index (CCFI) values by levels of factors in the simulation study design. These plots reproduce results shown in Figure 4 and extend them to wider ranges of values along each factor to examine performance under especially challenging conditions. Dotted vertical lines represent the rules of thumb for acceptable data parameters recommended by Meehl (1995).

“number of days used alcohol in the last 30 days” (for both items, 1 = *no days* to 7 = *30 days*), “age when first drank alcohol” and “age when first tried marijuana” (for both items, 1 = *never drank alcohol/tried marijuana* to 7 = *8 years old or younger*). The two sexual experience items were “age at first sexual intercourse” (1 = *never had sex* to 8 = *11 years old or younger*) and “number of sex partners” (1 = *never had sex* to 7 = *six or more people*).

A large-scale study by Romano, Tremblay, Vitaro, Zoccolillo, and Pagani (2001) estimated the prevalence rate of externalization disorders in adolescents to be 12.7%, so we used this base rate to assign cases to groups and estimate indicator validities and within-group correlations. The 12.7% of cases with the highest total externalization scores were assigned to the taxon and the other 87.3% to the complement. Seven of the eight indicators’ validity estimates were above Meehl’s (1995) recommended threshold of  $d = 1.25$ , as was the average for all eight indicators (mean  $d = 1.82$ ). The average within-group correlations among the eight indicators (mean  $r = .05$  for taxon,  $.19$  for complement) were well below Meehl’s recommended threshold of  $r = .30$ . Thus, these data appeared to be acceptable for an informative taxometric analysis.

Following the guidelines outlined above, we performed MAMBAC, MAXCOV, and L-Mode analyses just as in our simulation study, obtained CCFI values for each, and calculated the mean CCFI value. The rising MAMBAC and MAXCOV graphs with right-end cusps in Figure 7 did not match the prototypical

curve shapes expected for categorical data (peaked) or dimensional data (symmetrically concave for MAMBAC, flat for MAXCOV); though it was not symmetric, the L-Mode graph otherwise was similar to the unimodal prototype for dimensional data. An interpretation based on all three procedures’ curve shapes might suggest inconsistent or inconclusive results. In contrast, each procedure’s output for the empirical data exhibited greater concordance with that for dimensional than categorical comparison data. The CCFI values of  $.369$ ,  $.288$ , and  $.333$  obtained for MAMBAC, MAXCOV, and L-Mode, respectively, yielded a mean CCFI of  $.330$ , which fell below the range of ambiguous values defined using even broad dual thresholds ( $.40 \leq \text{CCFI} \leq .60$ ).<sup>4</sup> Revisiting Figure 3, we see that among the samples in our simulation study, a mean CCFI of  $.330$  was typical for dimensional data but exceptionally rare for categorical data. In fact, this set of CCFI values would have passed the most stringent test of consistency that we studied, the unanimity rule with broad dual thresholds. Consequently, the results of these analyses provide consistent support for

<sup>4</sup> The strength of these results is not attributable solely to the large sample size. When five randomly selected subsamples of  $n = 500$  cases apiece were analyzed in the same way, each of the mean CCFIs of  $.367$ ,  $.396$ ,  $.387$ ,  $.369$ , and  $.363$  was more consistent with dimensional than categorical structure by the stringent standard of broad dual thresholds.

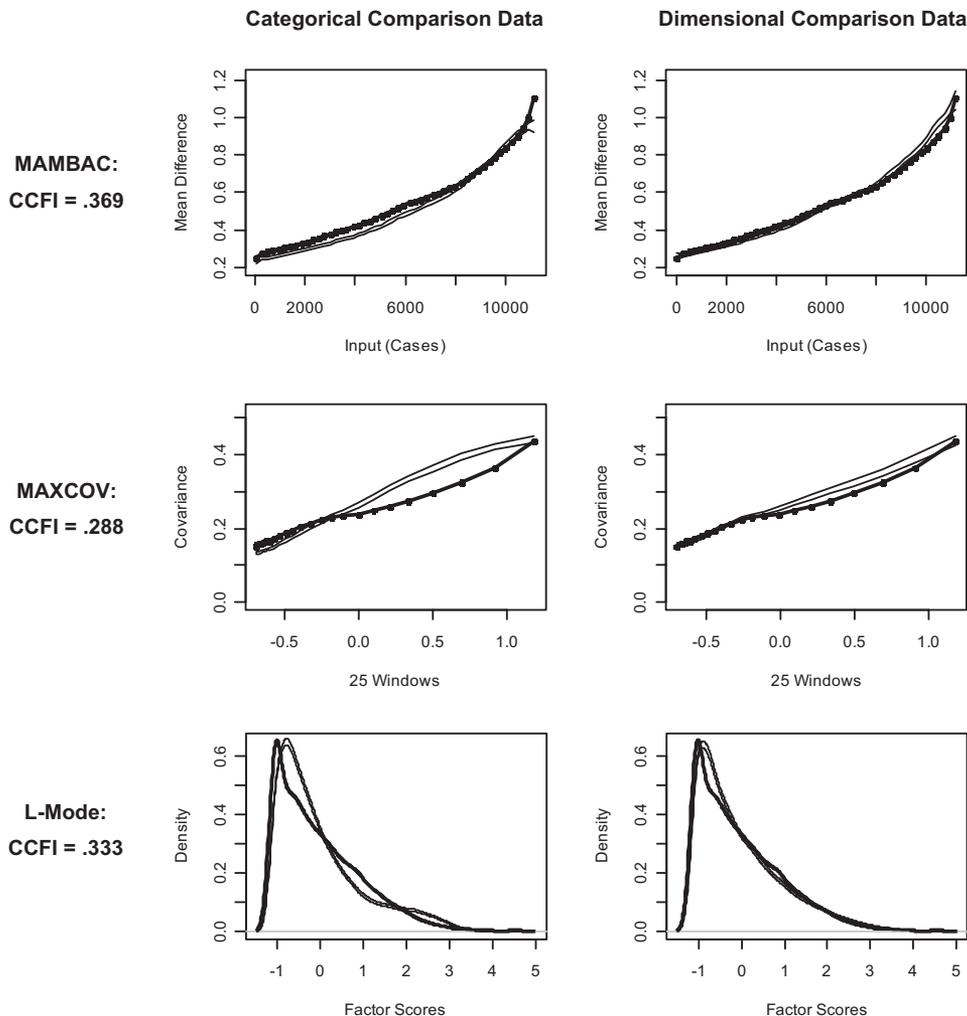


Figure 7. In each graph, results for the empirical data (dark curves) are superimposed above the band of results for comparison data; at each data point along the x-axis, the band depicts the mean of the y values for each sample of comparison data ( $\pm 1 SD$  of those values). The mean of the three CCFI values is .330. MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; L-Mode = latent mode; CCFI = comparison curve fit index.

the hypothesis that the latent structure of externalizing problems in adolescents is better characterized as dimensional than categorical.

### Discussion

The present study was designed to examine six related issues in the development, refinement, and evaluation of operationalizations of consistency testing. After we review the results to present tentative guidelines for an empirically supported way to practice consistency testing in taxometric analyses, we discuss some broader implications of this study.

First, for the same target data sets, how well does the CCFI for each of several taxometric procedures perform? Across the 100,000 target samples that spanned a wide range of data conditions, each taxometric procedure correctly distinguished categorical from dimensional data with an impressive accuracy rate. MAMBAC attained the greatest accuracy (96.3%), followed by L-Mode (94.7%) and then MAXCOV

or MAXEIG (92.3% for both). Expressed using the familiar metric of the area under a receiver operating characteristic curve,  $A = .998$ , .988, and .976, respectively, for these procedures. Provided that their data requirements are satisfied (see Meehl, 1995; Ruscio et al., 2006), each of these procedures should be considered for inclusion in a taxometric analysis.

Second, what are the consequences of using single versus dual thresholds for interpreting the CCFI? Even though the accuracy rates were relatively high for each procedure, they increased further when dual thresholds were used to set aside interpretationally ambiguous results. For example, using dual thresholds of .45 and .55, the accuracy rates for MAMBAC, L-Mode, and MAXCOV were 98.2%, 97.3%, and 95.8%, respectively. However, this comes at the cost of withholding judgment in 5.5%, 8.0%, and 10.5% of samples for MAMBAC, L-Mode, and MAXCOV, respectively. Using dual thresholds of .40 and .60 further increased accuracy

rates to 99.2%, 98.9%, and 97.9%, at the cost of withholding judgment in 12.4%, 18.7%, and 22.9% of samples, respectively. We recommend choosing thresholds prior to performing taxometric analyses, with the selection based on one's preferred balance between the risks of ambiguous and erroneous results. On the one hand, using a single CCFI threshold of .50 disallows ambiguous results and therefore guarantees that a conclusion of categorical or dimensional structure will be reached. On the other hand, the accuracy of conclusions can be increased by using dual thresholds. This increase comes at the price of setting aside more results as ambiguous, which would require withholding judgment more often. We cannot recommend a general-purpose threshold choice because this depends on how much risk of ambiguity one is willing to tolerate to achieve a gain in accuracy.

Third, does using multiple populations of categorical comparison data improve the performance of the CCFI? There is certainly no harm in using multiple populations of categorical comparison data, though the gain in accuracy was slight (less than 0.5% for each procedure studied here) and may not be worth the time and effort if one is performing a large number of analyses with a large sample of data.

Fourth, how should agreement across results from multiple taxometric procedures be defined? We compared the performance of a mean CCFI value as well as a unanimity rule and a majority rule for deciding between categorical, dimensional, and ambiguous outcomes. Holding constant the level of interpretable results, the mean CCFI achieved greater accuracy than the unanimity or majority rules. There appear to be no circumstances under which the majority rule can be recommended. One would have to be willing to dramatically increase the risk of ambiguous results to prefer the unanimity rule to the mean CCFI. For example, using dual thresholds of .40 and .60 achieved an accuracy of 99.85% for 86.0% of interpretable results with the mean CCFI, and an accuracy of 99.97% for 61.2% of interpretable results with the unanimity rule. It is difficult to conceive of circumstances under which one would be willing to absorb such a substantially increased risk of ambiguous results in exchange for such a slight gain in accuracy. This supports the use of the mean CCFI to combine results across multiple procedures.

Fifth, which taxometric procedures provide incremental validity? Correlations between the procedures' CCFI values showed that MAXCOV and MAXEIG yielded almost completely redundant results. We performed all subsequent analyses using MAXCOV or MAXEIG along with MAMBAC and L-Mode, and none of the results favored MAXCOV or MAXEIG by a nontrivial amount. For simplicity, we presented results only for MAXCOV, but we stress that they would have been nearly identical had we presented results only for MAXEIG. When procedures were used in combination, both MAMBAC and L-Mode always provided incremental validity by increasing accuracy rates relative to those for individual procedures or combinations of procedures from which they were excluded. MAXCOV did not always increase accuracy, and it appears to be the least important procedure to include for the purpose of consistency testing. Traditionally, MAXCOV has been dominant taxometric procedure, with MAXEIG taking its place in many studies in recent years (Ruscio et al., 2006). Even though we did not obtain evidence suggesting that consistency testing improves with the inclusion of MAXCOV or MAXEIG, it seems likely that readers or reviewers will continue to expect to see one of these procedures

performed in taxometric studies. In the present study, including MAXCOV or MAXEIG did no harm.

Sixth, how accurate can researchers expect results to be? We presented the probability that the mean CCFI was correct across its full range of values as well as its robustness across data conditions. As expected, the probability that CCFI values correctly identified categorical or dimensional data reached a minimum near .50 and was fairly symmetric about this value. Any mean CCFI value outside the dual thresholds of .45 and .55 attained a probability of being correct at or above .90. For mean CCFI values outside the dual thresholds of .40 and .60, the probability of being correct remained near 1.00.

At least as important is the finding that the excellent accuracy level of the mean CCFI was highly robust across data conditions within the limits for which taxometric analysis is recommended (Meehl, 1995). When conditions such as low indicator validity or high within-group correlations posed the greatest challenge to taxometric analysis, the percentage of ambiguous results increased but the percentage of incorrect results remained very low. Across all conditions studied, the absolute percentage of incorrect results never exceeded 3%, and the percent incorrect out of all interpretable results seldom exceeded 1%. Researchers can use the mean CCFI with confidence that they are unlikely to reach erroneous conclusions. This protection against erroneous conclusions weakened when data conditions extended beyond the recommended limits, underscoring the importance of taking Meehl's rules of thumb seriously.

On the basis of these results, we conclude that consistency testing in taxometric analysis can be operationalized successfully and implemented objectively. Recall that Meehl's goal was to sort results into three piles: strong evidence of categorical structure, strong evidence of dimensional structure, and ambiguous evidence suggesting that judgment should be withheld. The present results show that this was achieved by operationalizing consistency testing using the CCFI to quantify taxometric graphs. This approach relies on the parallel analysis of comparison data, which has been shown to work well for individual taxometric procedures. Combining the CCFI values for multiple procedures attained very high accuracy rates, and results that were not accurate were more often ambiguous than erroneous. Thus, consistency testing achieved its primary purpose, to set aside interpretationally ambiguous results and accurately distinguish categorical and dimensional data among the interpretable results.

Investigators have published taxometric studies featuring a diverse array of consistency tests, demonstrating that the practice itself has been implemented inconsistently. Focusing on the primary, graphical output of the taxometric procedures performed most often in published studies provides a template that would be easy for researchers to apply. This would reduce the idiosyncrasy in selecting consistency tests that has been observed in the taxometric literature, often giving rise to differences of opinion that could not be resolved empirically. For example, some investigators advocate an approach to consistency testing premised on the notion that if multiple estimates of certain structural model parameters (e.g., group base rates) agree within tolerable limits, this provides support for a categorical model. Unfortunately, it remains unclear what those tolerable limits might be. An approach based on the coherence of structural model parameter estimates may work

well, but this remains speculative because it has not been operationalized or studied rigorously.

In contrast, we present the first fully operationalized approach to consistency testing and subject it to extensive empirical trials involving a wide range of challenging data conditions. Our study places this approach on a solid empirical foundation, establishing for the first time empirically supported guidelines for the practice. We recommend performing MAMBAC, MAXCOV (or MAXEIG), and L-Mode analyses, obtaining the CCFI based on parallel analyses of categorical and dimensional comparison data for each, and calculating the mean of these three CCFI values. Dual thresholds of .45 and .55 seem to achieve a high accuracy rate among a large percentage of interpretable results—99.4% accuracy for 94.8% of samples in the present study—but there may be valid reasons to use more or less stringent thresholds. Provided that data conditions are within the recommended limits for taxometric analysis, this approach cannot guarantee accuracy but it does provide very strong protection against erroneous results by identifying as ambiguous most of the results that might otherwise have led to mistaken conclusions.

The present results provide empirical support for the CCFI method and we encourage researchers interested in alternative approaches to conduct similar Monte Carlo studies to assess the accuracy of these methods. We designed our study to cover a very wide range of data conditions, and we wrote our program such that the 100,000 target samples can be reproduced for study using other methods. Though we make no claim that our approach is optimal, we do note that it sets the bar pretty high. Investigators may or may not be able to operationalize consistency testing more effectively, but valuable insights may be gained in the process of attempting to do so. In the meantime, researchers can place a high degree of confidence in results obtained using the approach to consistency testing introduced here.

The present results hold implications for the data requirements of taxometric analysis as well as the possibility of performing exploratory taxometric research. As data conditions became more challenging, the percentage of ambiguous results increased. This underscores the importance of working with high-quality data (e.g., a large sample that contains as many valid and nonredundant indicators as possible) using well-validated instruments to assess the construct in question. The likelihood of obtaining ambiguous results increases to the extent that the data are ill-suited for the planned analyses. It appears that one can perform exploratory taxometric analyses with little risk of reaching mistaken conclusions. Opinions are mixed on the advisability of testing structural models in the absence of explicit, theory-based hypotheses. For example, Lenzenweger (2004) expressed reservations about taxometrics “undertaken with a minimum regard for a theory or a model” (p. 17), whereas Meehl (1999) outlined an exploratory method for scanning a large number of variables for clues to the structure of constructs that they represent. We take no firm position on the conditions under which one should or should not perform taxometric analyses in an exploratory mode. Our only contribution to this dialogue is to note that should one choose to engage in exploratory taxometric analyses, our operationalization of consistency testing should provide protection against mistaken conclusions provided that data meet the recommended standards for taxometrics. How one interprets the categorical, dimensional, or ambiguous results of the analysis goes beyond the implementation of the data-analytic procedures used. Even if hypotheses

were not derived theoretically, we agree with Lenzenweger and Meehl (2004) that theory should guide the interpretation of results when modeling latent variables.

We began this article by noting that consistency testing is a cornerstone of Meehl’s (1995) taxometric method. To realize the full potential of this model comparison technique requires that it be operationalized, including specification of the constituent procedures and decision rules for combining and interpreting their results. The present study is the first to rigorously examine an operationalized approach to taxometric consistency testing. This should prove helpful to investigators using the taxometric method, and we encourage others to operationalize and study alternative approaches to consistency testing.

## References

- 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., . . . Cox, B. (2007). Taxometric and factor analytic models of anxiety sensitivity: Integrating approaches to latent structural research. *Psychological Assessment*, *19*, 74–87.
- Broman-Fulks, J. J., Green, B. A., Berman, M. E., Olatunji, B. O., Arnau, R. C., Deacon, B. J., & Sawchuk, C. N. (2008). The latent structure of anxiety sensitivity—revisited. *Assessment*, *15*, 188–203.
- Broman-Fulks, J. J., Hill, R. W., & Green, B. A. (2008). Is perfectionism categorical or dimensional? A taxometric analysis. *Journal of Personality Assessment*, *90*, 481–490.
- Eaton, D. K., Kann, L., Kinchen, S., Shanklin, S., Ross, J., Hawkins, J., . . . Wechsler, H. (2008, June 6). *Morbidity and mortality weekly report (Vol. 57): Youth risk behavior surveillance—United States, 2007*. Retrieved from <http://www.cdc.gov/mmwr/preview/mmwrhtml/ss5704a1.htm>
- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). New York, NY: Oxford University Press.
- Franklin, L. C., Strong, D. R., & Greene, R. L. (2002). A taxometric analysis of the MMPI-2 Depression scales. *Journal of Personality Assessment*, *79*, 110–121.
- Grove, W. M. (2004). The MAXSLOPE taxometric procedure: Mathematical derivation, parameter estimation, consistency tests. *Psychological Reports*, *95*, 517–550.
- Haslam, N., & Kim, H. (2002). Categories and continua: A review of taxometric research. *Genetic, Social, and General Psychology Monographs*, *128*, 271–320.
- 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, NY: Wiley.
- Kotov, R., Schmidt, N. B., Lerew, D. R., Joiner, T. E., & Ialongo, N. S. (2005). Latent structure of anxiety: Taxometric exploration. *Psychological Assessment*, *17*, 369–374.
- Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology*, *114*, 537–550.
- Lenzenweger, M. F. (2004). Consideration of the challenges, complications, and pitfalls of taxometric analysis. *Journal of Abnormal Psychology*, *101*, 567–571.
- 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.

- McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.
- McLachlan, G., & Peel, D. (2001). *Finite mixture models*. New York, NY: Wiley-Interscience.
- Meehl, P. E. (1965). *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Report No. PR-65-2). Minneapolis, MN: University of Minnesota, Research Laboratories of the Department of Psychiatry.
- Meehl, P. E. (1968). *Detecting latent clinical taxa, II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Report No. PR-68-4). Minneapolis, MN: University of Minnesota, Research Laboratories of the Department of Psychiatry.
- 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. (1999). Clarifications about taxometric method. *Applied and Preventive Psychology*, *8*, 165–174.
- Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology*, *113*, 39–43.
- Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). New York, NY: Wiley.
- 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.
- 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.
- Parker, J. D. A., Keefer, K. V., Taylor, G. J., & Bagby, R. M. (2008). Latent structure of the alexithymia construct: A taxometric investigation. *Psychological Assessment*, *20*, 385–396.
- Romano, E., Tremblay, R. E., Vitaro, F., Zoccolillo, M., & Pagani, L. (2001). Prevalence of psychiatric diagnoses and the role of perceived impairment: Findings from an adolescent community sample. *Journal of Child Psychology and Psychiatry*, *42*, 451–461.
- 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. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*, 19–30.
- Ruscio, J. (2009). Assigning cases to groups using taxometric results: An empirical comparison of classification techniques. *Assessment*, *16*, 55–70.
- Ruscio, J., Brown, T. A., & Ruscio, A. M. (2009). A taxometric investigation of *DSM-IV* major depression in a large outpatient sample: Interpretable structural results depend on the mode of assessment. *Assessment*, *16*, 127–144.
- 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., & 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. (2009). Using comparison data to differentiate categorical and dimensional data by examining factor score distributions: Resolving the mode problem. *Psychological Assessment*, *21*, 578–594.
- 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.
- Strong, D. R., Greene, R. L., Hoppe, C., Johnston, T., & Olesen, N. (1999). Taxometric analysis of impression management and self-deception on the MMPI-2 in child-custody litigants. *Journal of Personality Assessment*, *73*, 1–18.
- Strong, D. R., Greene, R. L., & Kordinak, S. T. (2002). Taxometric analysis of impression management and self-deception in college student and personnel evaluation settings. *Journal of Personality Assessment*, *78*, 161–175.
- Strong, D. R., Greene, R. L., & Schinka, J. A. (2000). Taxometric analysis of MMPI-2 Infrequency Scales [*F* and *F(p)*] in clinical settings. *Psychological Assessment*, *12*, 166–173.
- Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage.
- Walters, G. D., Berry, D. T. R., Lanyon, R. I., & Murphy, M. P. (2009). Are exaggerated health complaints continuous or categorical? A taxometric analysis of the Health Problem Overstatement Scale. *Psychological Assessment*, *21*, 219–226.
- Walters, G. D., Brinkley, C. A., Magaletta, P. R., & Diamond, P. M. (2008). Taxometric analysis of the Levenson Self-Report Psychopathy Scale. *Journal of Personality Assessment*, *90*, 491–498.
- Walters, G. D., Diamond, P. M., Magaletta, P. R., Geyer, M. D., & Duncan, S. A. (2007). Taxometric analysis of the Antisocial Features Scale of the Personality Assessment Inventory in federal prison inmates. *Assessment*, *14*, 351–360.
- Walters, G. D., Duncan, S. A., & Mitchell-Perez, K. (2007). A taxometric investigation of the Psychopathy Checklist—Revised in a heterogeneous sample of male prison inmates. *Assessment*, *14*, 270–278.
- 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., Rogers, R., Berry, D. T. R., Miller, H. A., Duncan, S. A., McCusker, P. J., . . . Granacher, R. P., Jr. (2008). Malingering as a categorical or dimensional construct: The latent structure of feigned psychopathology as measured by the SIRS and MMPI-2. *Psychological Assessment*, *20*, 238–247.
- 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 July 1, 2009

Revision received October 27, 2009

Accepted October 28, 2009 ■