Using the Comparison Curve Fix Index (CCFI) in Taxometric Analyses:

Averaging Curves, Standard Errors, and CCFI Profiles

John Ruscio, Lauren M. Carney, Lindsay Dever,

Melissa Pliskin, and Shirley B. Wang

The College of New Jersey

Author Note

John Ruscio, Lauren M. Carney, Lindsay Dever, Melissa Pliskin, and Shirley B. Wang, Department of Psychology, The College of New Jersey.

Lauren M. Carney is now at the Department of Psychology, University of Connecticut. Lindsay Dever is now at the Department of Mathematics, Bryn Mawr College. Shirley B. Wang is now at the Department of Psychology, Harvard University.

After the lead author, all co-authors made comparably significant contributions to this paper and are listed in alphabetical order.

Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, Ewing, NJ, 08628. Contact: ruscio@tcnj.edu

Abstract

Determining whether a construct is more appropriately conceptualized and assessed in a

categorical or a dimensional manner has received considerable research attention in recent years.

There are a variety of statistical techniques to address this empirically, and Meehl's (1995)

taxometric method has been among the most widely-used methods applied to constructs in the

areas of personality and psychopathology. In taxometric analysis, the Comparison Curve Fit

Index (CCFI; Ruscio, Ruscio, & Meron, 2007) is an objective measure of whether parallel

analysis of categorical or dimensional comparison data better reproduce empirical data results.

The development and use of the CCFI helps to reduce the subjectivity involved in performing

taxometric analyses and interpreting the results. In a series of simulation studies, we examine the

use of the CCFI to flesh out some empirically supported guidelines. We find that a panel of

curves should be averaged to calculate a single CCFI value (rather than calculating the CCFI for

each curve and averaging these values), that an ambiguous range of CCFI values should be

defined using a fixed-width interval (rather than a multiple of the estimated standard error), and

that constructing a CCFI profile can help to differentiate categorical and dimensional data and

provide a less biased and more precise estimate of the taxon base rate than conventional

methods. Implications of these findings for taxometric research relevant to psychological

assessment are discussed along with ways to perform analyses consistent with these

recommendations.

Keywords: assessment; categories; dimensions; taxometric analysis; taxometric method;

comparison curve fit index; standard error; CCFI profile; taxon base rate; base rate estimation

Public Significance Statement

For the theory, research, and practice of psychological assessment, it is important to determine whether individuals should be classified into categories (e.g., psychopaths vs. non-psychopaths) or located along a continuous dimension (e.g., degree of psychopathy). The present research helps to refine data-analytic methods for determining as objectively as possible when to use categories and dimensions.

Using the Comparison Curve Fix Index (CCFI) in Taxometric Analyses:

Averaging Curves, Standard Errors, and CCFI Profiles

Psychologists, like other social and behavioral scientists, often wonder whether people differ on any given construct by belonging to discrete groups or by varying along a continuum. Researchers studying the assessment of constructs in the realm of personality and psychopathology have been especially interested in making this distinction (Ruscio, Walters, Marcus, & Kaczetow, 2010). For example, do depressed and non-depressed individuals form two separate groups of people, or does everyone fall along a continuous spectrum of depression? Either a categorical or a dimensional structural model may be more appropriate for any given construct (Meehl, 1995). Often, practitioners and researchers implicitly or explicitly endorse *a priori* preferences for how they will conceptualize and measure a construct. For instance, practitioners regularly diagnose mental disorders as either present or absent, treating individual differences as categorical in nature. On the other hand, researchers often evaluate disorders on a continuum of symptom severity, thus leaning towards a dimensional model.

Rather than relying on these preferences, empirically evaluating the structure of psychological constructs can have implications for assessment, classification, diagnosis, research design, statistical analysis, and theorizing (MacCallum, Zhang, Preacher, & Rucker, 2002; Meehl, 1992; Ruscio, Haslam, & Ruscio, 2006). For example, a measurement instrument intended to classify individuals into groups could use a relatively small number of items that maximize their discriminating power near the boundary between groups. In contrast, a measurement instrument intended to locate individuals along a continuum would require a larger number of items to make reliable and valid discriminations at all levels along that continuum (Ruscio & Ruscio, 2002). Thus, research should proceed quite differently depending on whether

the construct to be assessed is conceptualized as categorical or dimensional.

Meehl's (1995) taxometric method is designed to determine empirically whether it is more appropriate to use a categorical or dimensional model for a construct of interest, and it has been used in approximately 200 published studies (Haslam, Holland, & Kuppens, 2012, provide a quantitative review of the 177 studies available at the time). This method contains a family of data-analytic procedures that produce graphical results. Historically, the shapes of the curves in these panels of graphs were compared to those for prototypes produced from analyses of artificial categorical and dimensional data (Meehl & Yonce, 1994, 1996). Despite the clever and innovative methodology used to generate results, the subjective nature of the interpretive process imposed several limitations on taxometric research.

First, the categorical and dimensional curve prototypes were produced using a fairly small number of data sets with idealized data characteristics. Because the empirical data might differ in important ways (e.g., having skewed rather than normal score distributions), it was possible that comparisons to the prototypes could be misleading (Ruscio, Ruscio, & Keane, 2004). Second, the prototypes were produced using one particular implementation of each taxometric procedure. Many thoughtful decisions must be made to perform a taxometric procedure, and it was unknown how another implementation of the same procedure might affect the results (Ruscio et al., 2006). Third, the subjective process of interpreting taxometric results raised concerns about confirmation bias (Nickerson, 1998). Researchers might be tempted to see what they hoped or expected to see in the results, and it could be difficult to make a compelling case for the proper conclusions to draw. Fourth, the visual inspection of taxometric graphs placed severe limits on the size and scope of research on taxometric methodology. Investigating the most appropriate ways to perform taxometric analysis was difficult because it required that all

results in a simulation study be plotted and then interpreted by experts.

Reducing the Subjectivity of Taxometric Analysis

To reduce the subjectivity of taxometric analysis, and thereby to address each of the problems this entailed, Ruscio, Ruscio, and Meron (2007) introduced a technique to produce comparison graphs using parallel analyses of artificial categorical and dimensional data. These artificial data reproduced important characteristics of the empirical data (e.g., sample size, number of variables, marginal distributions, correlation matrices; Ruscio & Kaczetow, 2008), and they could be analyzed using the same procedural implementation as the empirical data. This yields taxometric graphs for data of known categorical and dimensional structure, holding everything else constant. Rather than relying on general-purpose prototypes, investigators could obtain comparison graphs tailored to the data and analysis plan in a particular study. This circumvented the first two problems, using a small number of prototypes for idealized data conditions that had been analyzed using just one procedural implementation.

The other two problems were addressed by developing the Comparison Curve Fit Index (CCFI). The CCFI is an objective measure of the extent to which the results for the empirical data are a closer match to those for the artificial categorical or dimensional comparison data. To calculate the CCFI, one begins by calculating the fit between graphs for the empirical data and the categorical comparison data. This is done via root mean square residual (RMSR) as follows:

$$\text{RMSR}_c = [\Sigma(y_e - y_c)^2 / N]^{1/2} \tag{1}$$

For each of the N data points on a taxometric graph, the residual is calculated as the distance between the point for the empirical data (y_e) and the corresponding point for the categorical comparison data (y_c). Perfect fit would yield residuals of 0, hence $\text{RMSR}_c = 0$. The larger the divergence between the points for the empirical and categorical comparison data, the larger the

value of $RMSR_c$. This calculation is performed a second time, using the points for the

dimensional rather than the categorical comparison data, to obtain $RMSR_d$. These two fit values

are then combined into the CCFI:

$$CCFI = RMSR_d / (RMSR_d + RMSR_c) \hspace{3cm} (2)$$

CCFI values range from 0 (strongest support for dimensional structure, obtained when

$RMSR_d = 0$ and $RMSR_c > 0$) to 1 (strongest support for categorical structure, obtained when

$RMSR_c = 0$ and $RMSR_d > 0$). A value of .50 is ambiguous (obtained when $RMSR_c = RMSR_d$).

A number of studies found that the CCFI effectively differentiates categorical and

dimensional data across a wide range of challenging data conditions, using a number of

taxometric procedures (Ruscio, 2007; Ruscio & Kaczetow, 2009; Ruscio & Marcus, 2007;

Ruscio et al., 2007; Ruscio & Walters, 2009, 2011; Ruscio et al., 2010; Walters, McGrath, &

Knight, 2010; Walters & Ruscio, 2009, 2010). Based not only on this evidence, but also on the

role that it played alongside other indicators of methodological strength in their quantitative

review of taxometric research, Haslam et al. (2012) concluded that "the most important historical

development in taxometric practice has been the analysis of simulated comparison data and use

of the CCFI … This technique has rapidly become dominant in taxometric research" (p. 911).

Using the CCFI reduces—but by no means eliminates—the threat of confirmation bias in

the interpretation of taxometric results. There are still concerns about researcher degrees of

freedom in the collection, preparation, analysis, and reporting of the data that ultimately produce

the CCFI values (Simmons, Nelson, & Simonsohn, 2011). Nonetheless, the degree of

subjectivity in the interpretive process itself is diminished with the use of an objective index.

In addition, the development of the CCFI enabled large-scale simulation studies of the

taxometric method. Because the results could be evaluated in an objective manner, the process

could be automated. This afforded the opportunity to generate and analyze many more samples

of data. The data could span wider ranges of values, which helped to establish the acceptable

data conditions for informative taxometric analyses, and be analyzed in multiple ways, which

helped to establish the most effective ways to implement taxometric procedures (see Ruscio et

al., 2011, for an overview). In addition, CCFI values obtained using multiple procedures could

be used to check the consistency of results (Ruscio et al., 2010).

<div align="center">**The Present Research**</div>

Although the CCFI has become ubiquitous in taxometric investigations, several important

questions remain unanswered. These are addressed through a series of simulation studies.

Averaging Curves vs. Averaging CCFI Values

What is the most appropriate way to calculate a single CCFI value when a taxometric

procedure yields a panel of curves? Because a cornerstone of the taxometric method is

consistency testing, procedures provide separate results by allowing variables to serve in the

required data-analytic roles in different configurations. For example, one procedure requires that

two variables serve in roles labeled as input and output. If there are three variables A, B, and C

available for analysis, these can be used in all possible input-output pairings—AB, BA, AC, CA,

BC, and CB—to yield a panel of 6 curves. From this panel of curves, a single CCFI value for the

procedure can be calculated in either of two different ways.

First, one could average the curves and then calculate the CCFI using this single,

averaged curve. Second, one could calculate a CCFI value for each curve and then average these

values. It is not obvious which method would work best, as both involve a type of aggregation

that should reduce the influence of sampling error on the final results. When the CCFI was

originally created, Ruscio et al. (2007) found it simpler to present and work with averaged

curves, and Ruscio (2016) found it easier to incorporate the curve-averaging approach into his

suite of taxometric programs. For the first time, we put this choice to an empirical test to

determine whether the CCFI-averaging approach should be used instead.

A secondary aim of this portion of our research was to determine whether CCFI values

calculated for individual curves within a full panel could be used to screen for particular

variables that are insufficiently valid to provide informative results.

Estimating the Standard Error of the CCFI

Like any statistic, the CCFI is subject to sampling error. In fact, there are two sources of

sampling error when calculating the CCFI. The first source of sampling error is, as usual, that

each unique sample of empirical data drawn from a particular population will yield somewhat

different statistical results. This source of sampling error can be reduced by using larger samples.

The second source of sampling error is introduced by the technique used to generate and analyze

the artificial comparison data. A finite but very large population of comparison data is generated,

and then random samples that match the size of the empirical data are selected for parallel

analysis. Both the generation of the population of artificial comparison data and the random

sampling of cases from this population contribute to sampling error in the CCFI. This can be

reduced by creating a larger population and performing parallel analyses of a larger number of

samples of comparison data.

Although there is no analytic formula to estimate the standard error (*SE*) of the CCFI, it

can be estimated as the standard deviation (*SD*) of an empirical sampling distribution (Efron &

Tibshirani, 1985). The first aim of this portion of our research was primarily descriptive, to

explore the estimated *SE* of the CCFI for multiple taxometric procedures performed using a wide

range of data conditions.

The second aim was to determine whether the estimated *SE* could be used to increase the

accuracy with which the CCFI can be used to differentiate between categorical and dimensional

data. To understand how this might help, recall that CCFI values of approximately .50 are

considered ambiguous. In their work on consistency testing, Ruscio et al. (2010) showed that

classifying data as categorical when CCFI > .50 and as dimensional when CCFI < .50 achieved

an accuracy rate of 98.0%. When setting aside as ambiguous the 5.2% of results with .45 ≤ CCFI

≤ .55, accuracy increased to 99.4%. Thus, an investigator willing to risk ambiguous findings can

reduce the risk of mistaken conclusions.

There are two ways to define the ambiguous range of CCFI values. The first is to use

fixed-width intervals centered around .50 (e.g., .50 ± .05). Alternatively, the ambiguous region

could be centered around .50 with a width calculated as a multiple of the estimated *SE* (e.g., .50

± 2 × *SE*). This would take sampling error into account in a manner analogous to constructing a

confidence interval and would situate the interpretation of CCFI values within a more

conventional statistical framework. It is not obvious which method would work best, so we put

this to an empirical test.

Constructing CCFI Profiles

When taxometric results appear to be categorical, one might want to estimate the relative

sizes of the two groups, referred to as the taxon and its complement. Each taxometric procedure

can be used to estimate the taxon base rate through a formula that is based on the shape of a

graph (Ruscio et al., 2006). While these formulas generally work well, they can be problematic.

Factors other than the actual taxon base rate can influence the shapes of taxometric graphs and

thereby bias base rate estimates (Ruscio et al., 2006).

Ruscio and Walters (2009) suggested an alternative method of base-rate estimation that

requires the construction of a profile of CCFI values. This is done by creating multiple populations of categorical comparison data that vary in their known base rates and analyzing many random samples from each to calculate a new $RMSR_c$ value, and thereby a new CCFI, at each taxon base rate. The CCFI profile is then constructed by plotting the CCFI value that was obtained for each population of categorical comparison data. The location of the peak value in the CCFI profile is used to infer the taxon base rate. Ruscio and Walters illustrated this method and found that as the taxon base rate in populations of categorical comparison data varied from P = .05 to .50 in increments of .05, the CCFI peaked at a value of .687 at P = .40. This was very close to the true taxon base rate of .416 in that sample.

The present research builds on that demonstration by systematically testing the bias and precision of taxon base rate estimates calculated using formulas and obtained using CCFI profiles. A secondary aim of this portion of our research is to determine whether a CCFI profile can help to differentiate categorical and dimensional data. If so, this would provide a method for doing so that does not require the user to speculate about the relative sizes of putative groups prior to analysis. This might be especially helpful in the study of constructs that are frequently assessed but that do not have explicit criteria (e.g., diagnostic algorithms) for assigning individuals to putative groups.

Simulation Methods

All simulations were performed in the R computing environment (R Core Team, 2017) using functions written by the authors. For each study, we generated a large number of samples of data that varied across a wide range of challenging but realistic conditions and analyzed these data using multiple taxometric procedures. The number of samples depended on the purpose of each study, how computing-intensive the necessary analyses would be, and the availability of

computing power at the time we performed that study. This research was exempt from review by

an instutitional review board because all data were simulated and none of the studies included

any human subjects.

Data Generation

When deciding on data conditions, we followed the lead of prior methodological research

in the taxometric literature (e.g., Ruscio & Kaczetow, 2009; Ruscio & Walters, 2009, 2011;

Ruscio et al., 2010; Walters & Ruscio, 2009). Unless otherwise noted, the following procedure

was used to generate samples for analysis.

For categorical samples, random values were drawn independently for the following

parameters: sample size (N = 300 to 1,000), number of variables (k = 3 to 8), taxon base rate (P =

.10 to .50), between-group validity (standardized mean difference between groups 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, or higher-

scoring group, relative to variance in its complement group).

Values of N, k, P, d, r, g, and h were drawn from uniform distributions (continuous for all

but N and k, which were 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 research, the smallest skew (γ_1) and

kurtosis (γ_2) values were $\gamma_1 = 0$, $\gamma_2 = 0$ for $g = 0$ and $h = 0$, respectively (a normal distribution) and the largest values were $\gamma_1 = 2.60$, $\gamma_2 = 38.89$ for $g = .30$ and $h = .15$, respectively; 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 for taxometric analysis.

To generate a categorical sample, the iterative technique of 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 identified as taxon members, with the remainder identified as members of the complement group. 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 groups was achieved by adding a constant to scores for taxon members such that the standardized mean difference equaled d.

For dimensional samples, data parameter values 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 (Meehl & Yonce, 1994). The iterative algorithm of 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 variables 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 samples with the intended correlations, distributions, and variance ratios. These data conditions go well beyond those used to produce the prototypical curve shapes that were used to interpret results prior to Ruscio et al. (2007).

Taxometric Procedures

We performed the procedures used most commonly in published taxometric studies. Below, we briefly describe each procedure and explain how it was implemented.

The MAMBAC (Mean Above Minus Below A Cut; Meehl & Yonce, 1994) procedure assigns one variable to the role of input and another to the role of output. Cases are sorted according to their scores on the input variable, 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 variable 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 variable. 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 input-output pairings.

The MAXEIG (MAXimum EIGenvalue; Waller & Meehl, 1998) procedure assigns one variable to the role of input and all remaining variables to the role of outputs.[1] Cases are sorted according to their scores on the input variable, 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 first (largest) eigenvalue of the covariance matrix for all output variables is calculated within each window. 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 plots the series of eigenvalues, on the ordinate, by the mean score for cases within each window along the input variable. For categorical data, the prototypical MAXEIG graph exhibits a peak near the score that optimally

separates the two groups. For dimensional data, the prototypical MAXEIG graph is flat rather than peaked. With $k \geq 3$ variables, one can perform a total of k MAXEIG analyses by assigning each variable to the role of input variable once.

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.

Calculating the CCFI

To calculate the CCFI, populations of artificial categorical and dimensional comparison data were generated and random samples were drawn from each for parallel analysis. For each taxometric procedure, the averaged curve across all samples of comparison data was used to calculate the $RMSR_c$ and $RMSR_d$ values, and from these a CCFI value was calculated.

To generate categorical comparison data, cases were assigned to groups using the base-rate classification method (Ruscio, 2009). This requires preliminary taxometric analysis to estimate the taxon base rate, and that was done using the MAMBAC and MAXEIG procedures. The mean base-rate estimate from those procedures was then used to determine what proportion of the highest-scoring cases (sorted by total scores across all variables) are assigned to the putative taxon, with the remainder assigned to the putative complement group. The base-rate classification method was used because it can be automated, it can be used with categorical and dimensional samples, and it provides a conservative estimate of CCFI accuracy.

Study 1a: Averaging Curves vs. Averaging CCFI Values

This study was designed to determine whether to average curves and calculate a single CCFI, or compute CCFIs for all curves in a panel of taxometric graphs and average these values.

Method

A total of 10,000 samples of data (5,000 categorical and 5,000 dimensional) were generated. For each sample, the MAMBAC and MAXEIG procedures were performed because each of them yields panels of curves (whereas L-Mode yields only a single curve). In addition to performing MAMBAC and MAXEIG as described above, additional methods of assigning variables to the input and output roles were used to generate panels containing differing numbers of curves. Specifically, MAMBAC was performed first by assigning variables to input-output roles in all possible pairings (which yields $k(k-1)$ curves for k variables), and then using the composite input variable method by which one variable serves as the output and all remaining variables are summed to form a composite input (which yields k curves). MAXEIG was performed first by assigning one variable to be the input with all others serving as outputs (which yields k curves), second using the composite input method (with two serving as outputs and the remainder summed to form the input, which yields $k(k-1)/2$ curves), and third using variables in all input-output-output triplets (which yields $k(k-1)(k-2)/2$ curves).

For each sample and each procedural implementation, the CCFI was calculated as usual from an averaged curve. Next, the CCFI was calculated separately for each curve in the panel and then these values were averaged. Accuracy was scored by determining the percentage of samples classified correctly as categorical or dimensional using a single CCFI threshold of .50.

Results and Discussion

For each procedural implementation, averaging curves yielded results that were as or more accurate than those for averaging CCFIs (see the top of Table 1). The largest difference

emerged when performing MAMBAC using variables in all possible pairings: Averaging curves classified 94.4% of the samples correctly, and averaging CCFIs classified 83.0% correctly. The only time that averaging curves and averaging CCFIs performed equally well was when performing MAXEIG using each variable as input (91.0% accuracy).

We examined the possibility that results might differ with more (or fewer) curves in each panel. After checking accuracy levels across varying numbers of curves, there was never an advantage to averaging CFFIs rather than averaging curves. A final series of analyses found that using trimmed means to average CCFIs—which would reduce the influence of any outliers due to anomalous curves—affected accuracy rates very little. No trimming proportion from .00 (the mean) to .50 (the median) yielded results superior to the averaged-curve CCFI.

In sum, these findings provide no reason to justify a change from the current practice of averaging curves to calculate a single CCFI. Regardless of how MAMBAC or MAXEIG was implemented, and regardless of whether means or trimmed means were used to aggregate CCFI values, calculating the CCFI after averaging curves in a panel always performed as well or better than calculating CCFI values for each curve and then averaging these values.

Study 1b: Identifying Insufficiently Valid Variables

Study 1a revealed that averaging curves yields equal or better results than averaging CFFIs. Study 1b was designed to determine whether calculating CCFIs for all curves can help to identify potentially problematic variables. Meehl's (1995) rule of thumb, supported by the findings of Ruscio et al. (2010), is that variables should possess a between-group validity of $d \geq$ 1.25 for taxometric analyses. If a researcher fails to estimate each variable's validity, or does so in a way that overestimates some values, calculating CCFIs for individual curves in a panel might be helpful to identify insufficiently valid variables that could weaken the analysis.

Method

To test this proposition, 10,000 samples of data (5,000 dimensional and 5,000 categorical) were generated as in Study 1a, with two exceptions. First, validity was substantially reduced either for one variable (for 5,000 samples) or two variables (for 5,000 samples). For categorical samples, the less-valid variables had their value of d cut in half. Because the original variables ranged from $d = 1.25$ to $d = 2.00$, less-valid variables' validities ($d = .62$ to 1.00) fell well below Meehl's (1995) threshold for taxometric analysis. For dimensional samples, the less-valid variables had their value of r_{xy} cut in half. Second, because MAXEIG requires at least three variables, the minimum number of variables in a sample was increased from three to four when one was less valid, and to five when two were less valid. To isolate the influence of individual variables on CCFI values, MAMBAC was implemented using variables in all possible pairings and MAXEIG was implemented using variables in all possible triplets.

Results and Discussion

In the first series of analyses, CCFI values were calculated by averaging curves. Surprisingly, dropping the less-valid variables actually reduced the accuracy with which categorical and dimensional data were distinguished (see the bottom of Table 1). Among categorical samples, removing the less-valid variables increased accuracy. With two less-valid variables, accuracy increased from 94.9% with all variables to 98.5% after removing the two less-valid variables. Among dimensional samples, on the other hand, accuracy decreased when the less-valid variables were removed. With two less-valid variables, accuracy dropped from 92.7% with all variables to 87.0% after removing the two less-valid variables. Overall, though, gains in accuracy for categorical samples were more than offset by reductions in accuracy for dimensional samples. Because taxometric analysis is performed to infer structure, arguably this

total accuracy rate is most important.

The preceding results involved removing variables known to be less valid. In practice, one would need a procedure for identifying less-valid variables. No such procedure would itself function flawlessly. Thus, our findings represent an upper limit to how effectively CCFI values for individual curves might serve in the identification and removal of less-valid variables. In short, we found no evidence that would support the calculation of CCFI values for all curves in a panel to help check the validity of each variable. Please note that these findings do not justify the inclusion of variables that researchers can in fact identify as insufficiently valid for taxometric analysis. We only created less-valid variables to determine whether CCFI values could be used to identify them for removal. A negative answer to this question in no way implies that other steps should not be taken to identify and remove insufficiently valid variables. In particular, we strongly recommend assigning cases to putative groups to check each variable's validity prior to performing any taxometric analyses.

Study 2: Estimating the SE of the CCFI

This study was designed to examine the sampling error in the CCFI and to determine whether an estimate of the *SE* of the CCFI would be useful to define an ambiguous range of CCFI values following a procedure analogous to constructing a confidence interval.

Method

A total of 1,000 samples of data (500 categorical and 500 dimensional) were generated. To estimate *SE*s, we treated each sample as a population and bootstrapped 25 samples in order to calculate the CCFI for each of these bootstrapped samples. Efron and Tibshirani (1993) suggest a minimum of 25 bootstrap samples for statistical applications, and this seemed sufficient because taxometric research requires large samples of data that should yield relatively small *SE*s. The *SE*

of the CCFI was estimated for each target data set as the *SD* of the distribution of CCFIs across the 25 bootstrap samples.

Results and Discussion

The estimated *SE* was generally low, but it varied modestly across the MAMBAC (*M* = .064, *SD* = .030), MAXEIG (*M* = .061, *SD* = .018), and L-Mode (*M* = .052, *SD* = .020) procedures. The estimated *SE* of the mean CCFI across these three procedures was fairly low for most samples (*M* = .044, *SD* = .014). It was also a bit smaller for categorical samples (*M* = .037, *SD* = .012) than for dimensional samples (*M* = .050, *SD* = .012). As might be expected from the large-sample nature of taxometric research, the CCFI exhibits relatively small amounts of sampling error, particularly the mean CCFI for multiple taxometric procedures.

Although having some sense for how much sampling error typically accompanies a CCFI value, the chief aim of this study was to determine whether it would be useful to construct an ambiguous region for the CCFI using multiples of the estimated *SE* rather than as fixed-width intervals. To test this, ambiguous regions were constructed using fixed-width intervals ranging from .50 ± 0.00 to .50 ± 0.10 as well as intervals ranging from .50 ± *SE* × 0.00 to .50 ± *SE* × 2.00. For each region, the percentage of ambiguous results was recorded and then accuracy was calculated as the percentage of samples correctly classified after setting aside the ambiguous ones. Figure 1 (top) shows that the accuracy-ambiguity trade-off was no more favorable when constructing the ambiguous region using multiples of the estimated *SE* than when using fixed-width intervals. For a given amount of ambiguity, accuracy was comparable or better using fixed-width intervals.

Because we were concerned that a ceiling effect on accuracy might have masked a difference between these two methods, we generated an additional 10,000 samples of data (5,000

categorical and 5,000 dimensional) using considerably more challenging data conditions.

Parameters for categorical samples in these more challenging data were randomly drawn from

the following: sample size (N = 100 to 800), number of variables (k = 3 to 6), taxon base rate (P

= .05 to .25), between-group validity (d = 1.00 to 1.75), within-group correlation (r = .10 to .30),

asymmetry (g = .10 to .30), tail weight (h = .05 to .15), and variance ratio (VR = 2 to 4). Many of

these samples should be extremely challenging for taxometric analysis because one or more data

parameters will be near or below the recommended thresholds. Figure 1 (bottom) shows that for

a given amount of ambiguity, accuracy was superior when the ambiguous region was defined by

fixed-width intervals rather than multiples of the estimated SE.

Given these findings, we do not recommend that researchers follow the computing-

intensive process required to estimate the SE of the CCFI. We found no benefit when using this

value to construct an ambiguous region of CCFI values in a manner analogous to the

construction of a confidence interval. This places the practice of using fixed-width intervals to

define an ambiguous region of CCFI values on a foundation of empirical support.

Study 3: Constructing CCFI Profiles

This study was designed to determine whether a CCFI profile, constructed by calculating

a series of CCFI values for populations of categorical comparison data with systematically

varying taxon base rates, could improve base rate estimates or help to differentiate categorical

and dimensional data.

Method

A total of 100,000 samples (50,000 categorical and 50,000 dimensional) were generated;

these were identical to the samples created by Ruscio et al. (2010). Each was analyzed using the

MAMBAC, MAXEIG, and L-Mode procedures. CCFI profiles were constructed by obtaining

separate CCFI values for populations of categorical comparison data with taxon base rates

spanning the range of .025 to .975 in increments of .025, for a total of 39 CCFI values. Each

procedure yielded a CCFI profile, and taking the mean CCFI at each of the 39 base rates yielded

an additional profile.

Base rate estimates were calculated using formulas provided for MAMBAC (Meehl &

Yonce, 1994), MAXEIG (Ruscio et al., 2006, adapted from the technique for the MAXCOV

procedure in Meehl & Yonce, 1996), and L-Mode (Waller & Meehl, 1998). The base rate was

also estimated using each CCFI profile. First, the profile was smoothed. Trial and error revealed

that a 6th degree polynomial worked well. Then, the location of the peak in this smoothed curve

served as an estimate of the base rate. See Figure 2 for an illustration of using CCFI profiles to

estimate the base rate.

In addition to using CCFI profiles to estimate base rates, we tested five methods for using

CCFI profiles to differentiate categorical and dimensional data. The first method was to choose

one data point—one CCFI value—at random from the profile. This serves as a baseline against

which to judge the accuracy of other methods. The second method was to choose the most

extreme CCFI value, meaning the value most distant from .50. The expectation would be that the

most extreme data point on a CCFI profile would be well above .50 for categorical data and well

below .50 for dimensional data. The third method was to calculate the mean of all CCFI values

on the profile, which would reduce the influence of sampling error. The final two methods also

incorporate all CCFI values in the profile, but they attempt to give greater weight to those that

might be more useful. The fourth method calculates weights as the (vertical) distance from each

data point to .50, which reduces the weight assigned to ambiguous CCFI values. The fifth

method calculates weights as a function of the (horizontal) distance from each data point to the

estimated base rate. Specifically, the weights were calculated as x / P' for data points with x

values at or below the estimated base rate P' and $(1 - x) / (1 - P')$ for data points with x values

above the estimated base rate. This yields a triangular pattern of weights (see Figure 3 for an

example). For each of these five methods, accuracy was calculated as the percentage of samples

correctly classified as categorical or dimensional using a single CCFI threshold at .50.

Results

      We assessed the accuracy of base rate estimates using residuals (estimated – actual value)

as a measure of bias and absolute residuals as a measure of precision. The results for the 50,000

samples of categorical data are shown in Figure 3. The formula for MAMBAC yielded

substantial overestimates of the base rate, as shown in the very poor results for bias and

precision. Using a CCFI profile for MAMBAC performed considerably better in terms of bias

and precision. The formulas for MAXEIG and L-Mode yielded much less biased and more

precise estimates than did the formula for MAMBAC. Using CCFI profiles performed a bit

worse than the formula for MAXEIG and a bit better than the formula for L-Mode.

      Taking a mean of the estimates from formulas for all three procedures fared poorly

because of the inclusion of the MAMBAC results. In sharp contrast, using the mean CCFI profile

provided estimates with the least bias (*Mdn* = .011, *IQR* = -.002 to .026) and greatest precision

(*Mdn* = .017, *IQR* = .008 to .031) of any method in this study.

      Next, the accuracy with which categorical and dimensional data were differentiated using

CCFI profiles was examined. The results for all five methods are shown in Table 2. Perhaps the

most surprising finding is how accurate the randomly chosen CCFI values were, particularly for

the MAMBAC and mean CCFI profiles. This suggests a remarkable robustness in the CCFI to

the misspecification of the base rate when creating populations of categorical comparison data.

An accuracy rate of 97.0% was attained by randomly selecting a base rate anywhere from .025 to .975 and calculating the mean of the CCFI values for MAMBAC, MAXEIG, and L-Mode using this single, completely arbitrary population of categorical comparison data. This is certainly not a practice we recommend, but it does set a high bar for improving the results by using the information contained in a CCFI profile more effectively.

Despite this strong start, the other methods did perform better. Using the most extreme CCFI value was a substantial improvement for MAMBAC, MAXEIG, and L-Mode CCFI profiles. The best results for MAMBAC CCFI profiles were obtained using the mean of all CCFI values or the weighted mean based on the vertical distance of each data point to .50. The best results for MAXEIG and L-Mode CCFI profiles were obtained using the weighted mean based on the horizontal distance of each data point to the base rate estimate. This method yielded the best results in this study—98.2% accuracy—when applied to the mean CCFI profiles. This compares favorably with all of the methods tested by Ruscio et al. (2010) using the same 100,000 samples.

In Ruscio et al.'s (2010) study of consistency testing, the bottom-line recommendation was to calculate the mean CCFI and interpret it using dual thresholds that treat $.45 \leq \text{CCFI} \leq .55$ as ambiguous. Doing so yielded ambiguous results for 5.2% of the samples and attained an accuracy of 99.4% for the remaining samples. Based on results from the present study, it appears that aggregate CCFIs using the weighted mean based on the horizontal distance to the base rate estimate can be interpreted similarly. Applying the same dual thresholds yielded ambiguous results for 5.7% of the samples and attained an accuracy of 99.5% for the remaining samples.

Discussion

Constructing a CCFI profile is a computing-intensive process, requiring the generation of

many populations of comparison data and the parallel analysis of a very large number of samples

drawn from all of these populations. This may be worth the trouble, however, when one wishes

to estimate the taxon base rate for data believed to be categorical. Using the mean CCFI profile

provided estimates that were considerably less biased and more precise than those obtained using

the conventional formula for any procedure or the average of the results across procedures.

Using the mean CCFI profile also attained a high level of accuracy in differentiating categorical

and dimensional data. When selecting a single data point at random, accuracy was 97.0%, and

when all data points were weighted based on their distance to the base rate estimate, accuracy

was 98.2%. Likewise, interpreting these CCFIs using dual thresholds that treat the intermediate

values as ambiguous can improve validity further after setting aside a modest proportion of

findings as ambiguous.

General Discussion

One of the fundamental issues in psychological assessment is whether the goal is to

assign individuals to groups or to locate their positions along a continuum. Theory, research, and

practice in assessment depends critically on the relative fit of categorical and dimensional

models obtained in empirical studies (Meehl, 1992; Ruscio & Ruscio, 2002). Meehl's (1995)

taxometric method is one empirically supported tool for studying the latent structure of

psychological constructs, and the present research helps to minimize the subjectivity of

performing taxometric investigations and interpreting their findings.

The use of the CCFI in taxometric analysis has become mainstream, with 37 taxometric

studies published from 2011 to 2016 all using CCFI values to distinguish between categorical

and dimensional data. The current trio of studies answered several lingering questions about the

methods used to calculate CCFI values, the extent and usefulness of its *SE*, and the use of CCFI

profiles to estimate base rates and differentiate categorical and dimensional data. Results of the

first two simulation studies largely support extant guidelines and methods for calculating and

interpreting CCFI values, thereby providing empirical support for previously untested practices.

The results from Study 3, on the other hand, demonstrate the utility of a new, computing-

intensive procedure for constructing CCFI profiles.

Studies 1a and 1b provided no support for either of two justifications for calculating

CCFIs for all curves in a full panel. Even if there are one or two less-valid variables in the

analysis—so much so that they fall well below the acceptable threshold for taxometric

analysis—it appears that researchers should still use the conventional method of averaging

curves to calculate the CCFI using all available data. Though there is surely some level of

validity below which weak variables would jeopardize taxometric results, it does not appear that

calculating CCFI values for individual curves is a useful way to identify them. Any variable

whose validity is that low should be identifiable in the usual manner, by estimating d across

cases assigned to putative groups prior to performing any taxometric analyses.

Study 2 provided no evidence to support the calculation and use of the *SE* of the CCFI in

defining ambiguous regions of CCFI. This confirms the effectiveness of setting ambiguous

regions using fixed-width intervals. Although setting intervals using the *SE* of the CCFI does not

seem to improve the favorability of the accuracy-ambiguity trade-off, researchers must still

decide whether and how far to extend an ambiguous region of CCFI values. Study 2 also

revealed that the *SE* of the CCFI was fairly small for all three taxometric procedures, and even

smaller when using a mean CCFI value calculated using multiple procedures.

Study 3, on the other hand, broke new ground by demonstrating two reasons that can

warrant the computing-intensive process required to construct a CCFI profile. The first reason is

that estimating base rates using CCFI profiles outperformed the conventional base rate

estimation formulas, both in terms of lower bias and greater precision. In particular, a smoothed

profile of mean CCFI values seems to work well. When the results of taxometric analysis suggest

that individuals should be classified into groups, determining their relative sizes is an extremely

important first step for doing so as accurately as possible. Whether using Bayes' theorem, the

base-rate classification method (Ruscio, 2009), or another technique, the taxon base rate is a

crucial piece of information. Using a CCFI profile is a promising way to estimate this as well as

possible from taxometric results.

The second reason to consider constructing a CCFI profile is that using a weighted mean

of all CCFI points along a profile (based on the horizontal distance of each point to the estimated

base rate) differentiated between categorical and dimensional data with slightly greater accuracy

than any alternative techniques applied to the same samples in prior research. Interestingly, even

choosing a value at random from a CCFI profile achieved an accuracy rate of 97%. Though not a

method to recommend, the surprisingly high accuracy of this arbitrary method demonstrates that

CCFI values are quite robust to misspecifications of the taxon base rate when generating

populations of categorical comparison data. Whether as a supplement to or a replacement for

other methods, using a CCFI profile can help to determine the relative fit of categorical and

dimensional models in a taxometric study. The present findings support the use of an aggregate

CCFI that can be interpreted using a single threshold at .50 or dual thresholds that set aside

ambiguous results (e.g., $.45 \leq CCFI \leq .55$) to boost validity for non-ambiguous values.

CCFI profiles might be especially useful when prior theory and research provide no clear

method for assigning cases to putative groups that can be used to generate a population of

categorical comparison data. Constructing a CCFI profile requires no such information. At the

same time, if prior theory and research do provide a method for assigning cases to putative groups, this method should be used rather than the base-rate classification technique. For example, to test the structure of a construct identified using a set of diagnostic criteria, these criteria should be used to assign cases to putative groups rather than using a series of taxon base rate estimates to construct a CCFI profile. The latter approach would be needlessly computing intensive and none of the populations of categorical comparison data might adequately represent the construct identified by the diagnostic criteria.

In recent years, most researchers performing taxometric analysis have been using free R code developed and updated by Ruscio (2016). In part because this suite of programs did not allow the construction of CCFI profiles, it has been replaced by an R package for taxometric analysis that is now available at https://cran.r-project.org/package=RTaxometrics (Wang & Ruscio, 2017). The modular functions in the RTaxometrics package follow contemporary R style for programming and documentation and run more efficiently than the programs they replace. Similarly, the new functions are grounded firmly in the empirical literature on the implementation of taxometric analysis. Both the default and the optional ways to perform taxometric procedures were established with regard for the findings of research on taxometric methodology. The new functions also are designed to be as user-friendly as possible when entering commands and working with the output. We anticipate that researchers will find the RTaxometrics package useful to perform taxometric analysis in ways consistent with the empirical literature on best practices, including but not limited to the present set of simulation studies. The most significant contribution of the present research to this endeavor is that it motivated the development of a new function to construct CCFI profiles. Just as the CCFI itself fairly quickly became standard practice in taxometric research, we expect the use of CCFI

profiles to become an attractive analytic tool for researchers studying the latent structure of

constructs relevant to the assessment of personality and psychopathology.

References

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of*

> *Psychology, 28*, 97–104.

Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy.

> *Behaviormetrika, 17*, 1-35. doi:10.2333/bhmk.12.17_1

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

> & Hall.

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and

> psychopathology: A quantitative review of taxometric research. *Psychological*
>
> *Medicine, 42*, 903-920. doi:10.1017/S0033291711001966

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.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of

> dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40.
>
> doi:10.1037/1082-989X.7.1.19

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in

> kind. *Journal of Personality, 60*, 117-174. doi:10.1111/j.1467-6494.1992.tb00269.x

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in

> psychopathology. *American Psychologist, 50*, 266-275. doi:10.1037/0003-066X.50.4.266

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*(3, Pt 2), 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*(3, Pt 2), 1091-1227.

    doi:10.2466/pr0.1996.78.3c.1091

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures.

    Psychological Bulletin, 105, 156–166.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review*

    Of General Psychology, *2*, 175-220. doi:10.1037/1089-2680.2.2.175

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for

    Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ruscio, J. (2007). Taxometric analysis: An empirically-grounded approach to implementing the

    model. *Criminal Justice and Behavior, 34*, 1588-1622. doi:10.1177/0093854807307027

Ruscio, J. (2009). Assigning cases to groups using taxometric results: An empirical comparison

    of classification techniques. *Assessment*, *16*, 55-70. doi:10.1177/1073191108320193

Ruscio, J. (2016). *Taxometric programs for the R computing environment: User's manual*.

    Available on the world wide web at http://ruscio.pages.tcnj.edu/quantitative-methods-

    program-code/

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method*: *A*

    practical guide. Mahwah, NJ: Lawrence Erlbaum Associates.

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative

    algorithm. *Multivariate Behavioral Research*, *43*, 355-381.

    doi:10.1080/00273170802285693

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the

robustness of taxometric analysis. *Multivariate Behavioral Research, 44*, 259-280.

   doi:10.1080/00273170902794248

Ruscio, J., & Ruscio, A. M. (2008). Categories and dimensions: Advancing psychological

   science through the study of latent structure. *Current Directions In Psychological

   Science, 17*, 203-207. doi:10.1111/j.1467-8721.2008.00575.x

Ruscio, J., Ruscio, A. M., & Carney, L. M. (2011). Performing taxometric analysis to distinguish

   categorical and dimensional variables. *Journal Of Experimental Psychopathology, 2*,

   170-196. doi:10.5127/jep.010910

Ruscio, J., & Ruscio, A. M., & Keane, T. M. (2004). Using taxometric analysis to distinguish a

   small latent taxon from a latent dimension with positively skewed indicators: The case of

   Involuntary Defeat Syndrome. *Journal of Abnormal Psychology, 113*, 145-154.

   doi:10.1037/0021-843X.113.1.145

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. doi:10.1080/00273170701360795

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. doi:10.1037/a0016558

Ruscio, J., & Walters, G. D. (2011). Differentiating categorical and dimensional data with

   taxometric analysis: Are two variables better than none? *Psychological Assessment, 23*,

   287-299. doi:10.1037/a0022054

Ruscio, J., Walters, G. D., Marcus, D. K., & Kaczetow, W. (2010). Comparing the relative fit of

   categorical and dimensional latent variable models using consistency tests. *Psychological

Assessment, 22, 5-21. doi:10.1037/a0018259

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed

flexibility in data collection and analysis allows presenting anything as

significant. *Psychological Science*, *22*, 1359-1366. doi:10.1177/0956797611417632

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types

from continua*. Thousand Oaks, CA: Sage.

Walters, G. D., McGrath, R. E., & Knight, R. A. (2010). Taxometrics, polytomous constructs,

and the comparison curve fit index: A Monte Carlo study. *Psychological Assessment*, *22*,

149-156. doi:10.1037/a0017819

Walters, G. D., & Ruscio, J. (2009). To sum or not to sum: Taxometric analysis with ordered

categorical assessment items. *Psychological Assessment*, *21*, 99-111.

doi:10.1037/a0015010

Walters, G. D., & Ruscio, J. (2010). Where do we draw the line? Assigning cases to subsamples

for MAMBAC, MAXCOV, and MAXEIG taxometric analyses. *Assessment*, *17*, 321-333.

doi:10.1177/1073191109356539

Wang, S. B., & Ruscio, J. (2017). RTaxometrics: An R package for taxometric analysis.

Manuscript submitted for publication.

Footnote

[1]Because the MAXCOV (MAXimum COVariance; Meehl & Yonce, 1996) procedure

operates in very similar ways and provides very similar results, we chose not to use both

procedures. The fundamental difference is that whereas MAXCOV involves only two output

indicators so that their covariance can be calculated within each ordered subsample of cases,

MAXEIG can accommodate more than two output indicators by calculating the first (largest)

eigenvalue of their covariance matrix (a variance-covariance matrix modified by replacing each

variance with 0). In a study of how best to implement taxometric procedures, Walters and Ruscio

(2010) found that both MAXCOV and MAXEIG worked best when using overlapping windows

rather than nonoverlapping intervals and when using a small, fixed number of windows. At 25

windows, the results for MAXEIG and MAXCOV were nearly identical, with a very slight edge

for MAXEIG. In their study of consistency testing, Ruscio et al. (2010) performed MAXCOV

and MAXEIG analyses with 50 windows on the same 100,000 samples included in the present

Study 3 and found that the CCFIs for these procedures correlated at $r = .999$. Using a single

threshold at CCFI = .50, MAXCOV achieved an accuracy of 92.34% and MAXEIG achieved

92.28%. Using the RTaxometrics package described in the General Discussion, only the

MAXEIG procedure is included. However, the default implementation assigns variables to input-

output indicator roles in all possible triplets. Thus, within each ordered subsample of cases the

eigenvalue will equal the absolute value of the covariance between these two indicators. Because

it is highly unusual for covariances to be negative in taxometric analyses, and it is virtually

impossible for this to occur in an interpretively important region of a curve, the results will be

identical or nearly identical to what would be obtained using MAXCOV.

Table 1. *Percent of Samples Correctly Classified, Studies 1a and 1b*

Study 1a		
Procedure and Implementation	Averaging Curves	Averaging CCFIs
MAMBAC, all input-output pairings	94.4	83.0
MAMBAC, composite input variable	93.8	93.6
MAXEIG, all input-output-output triplets	91.0	87.5
MAXEIG, single input variable	91.0	91.0
MAXEIG, composite input variable	84.5	83.5

Study 1b			
Variables In Analysis	All Samples	Categorical Samples	Dimensional Samples
5,000 samples with 1 less-valid variable			
All variables	94.8	96.2	93.4
All but 1 less valid variable	94.7	98.3	91.2
5,000 samples with 2 less-valid variables			
All variables	93.8	94.9	92.7
All but 2 less valid variables	92.8	98.5	87.0

Notes. CCFI = Comparison Curve Fit Index; MAMBAC = Mean Above Minus Below A Cut; MAXEIG = MAXimum EIGenvalue.

**Table 2.** *Percent of Samples Correctly Classified, Study 3*

| | CCFI Profile | | | |
Method	MAMBAC	MAXEIG	L-MODE	Mean
Random CCFI	92.5	74.2	75.4	97.0
Extreme CCFI	95.3	91.8	90.9	97.2
Mean CCFI	97.4	89.9	92.2	97.3
Weighted mean (vertical)	97.2	91.4	91.9	97.3
Weighted mean (horizontal)	95.8	92.4	94.6	98.2

Notes. CCFI = Comparison Curve Fit Index; MAMBAC = Mean Above Minus Below A Cut; MAXEIG = MAXimum EIGenvalue; L-Mode = Latent Mode.
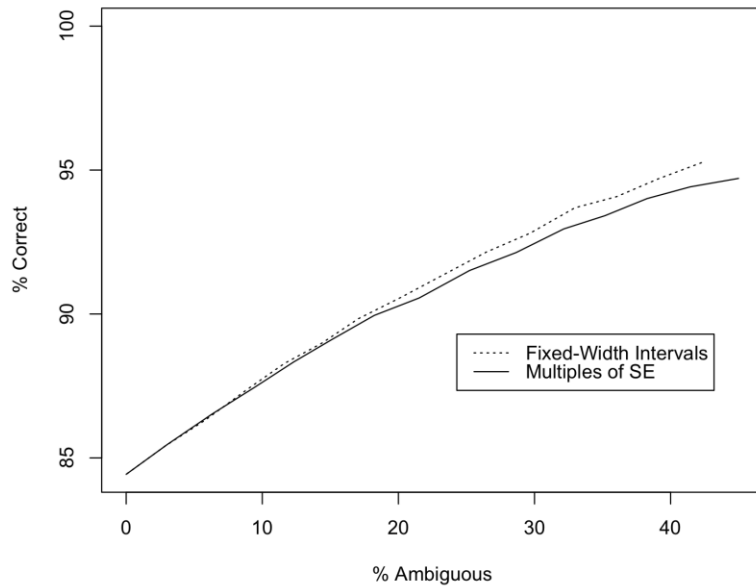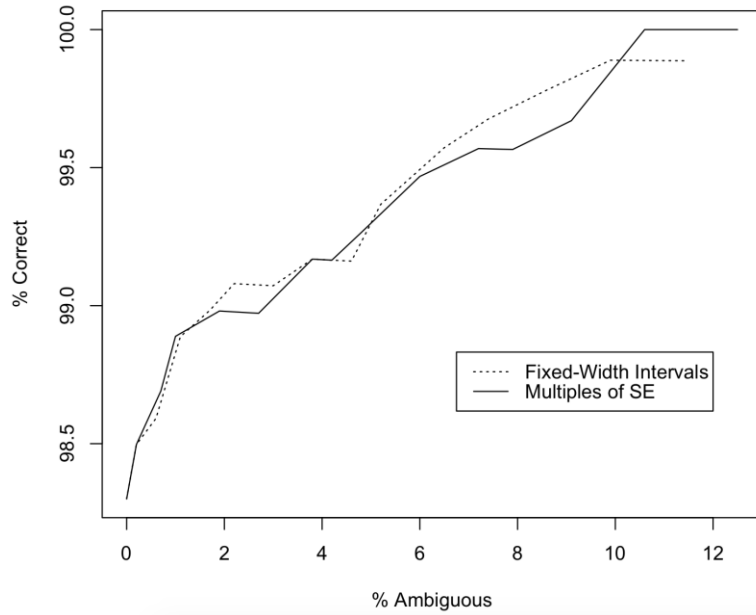
Figure 1. Accuracy-ambiguity trade-off, Study 2. The top graph shows results for the first 1,000 samples, and the bottom graph shows results for the 10,000 additional samples that were much more challenging. SE = standard error.
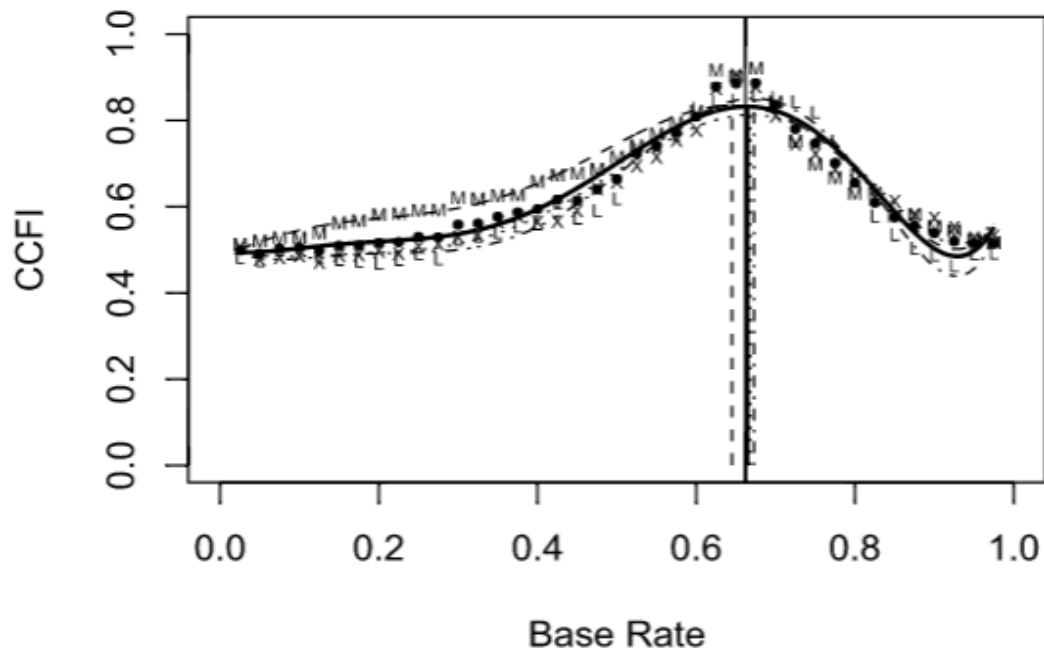
Figure 2. Comparison Curve Fit Index (CCFI) values plotted for analyses using 39 populations of categorical comparison data that varied across taxon base rates of $P = .025$ to .975 in increments of .025. CCFI profiles are labeled as M for MAMBAC (Mean Above Minus Below A Cut), X for MAXEIG (MAXimum EIGenvalue), and L for L-Mode (Latent Mode). Solid data points represent the mean CCFI profile. A vertical line is plotted from the peak of each smoothed curve to the corresponding base rate estimate on the x axis. The actual base rate is plotted as a solid vertical line.
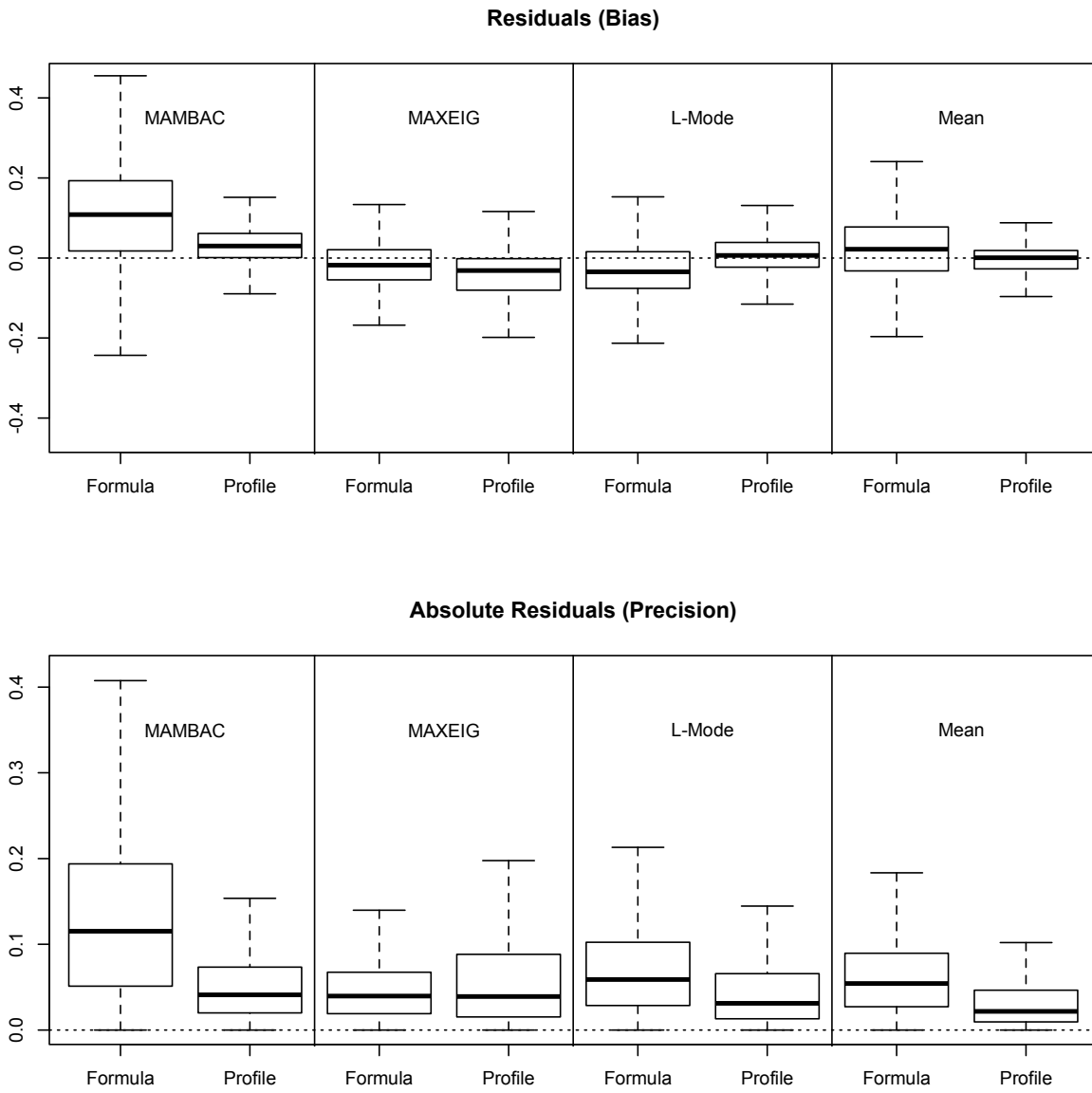
Figure 3. Accuracy of base-rate estimates using formulas and CCFI profiles, Study 3. MAMBAC = Mean Above Minus Below A Cut; MAXEIG = MAXimum EIGenvalue; L-Mode = Latent Mode.
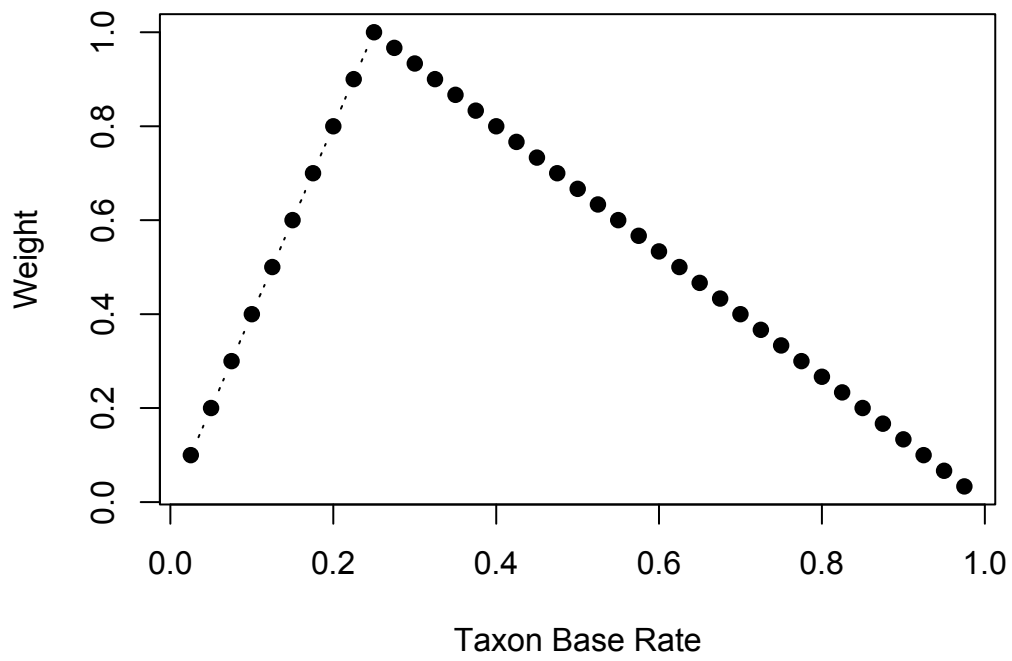
Figure 4. Example of horizontal weights assigned to each data point on a Comparison Curve Fit Index (CCFI) profile if the taxon base rate estimate is .25.