Assessment
17(3) 321–333

DOI: 10.1177/1073191109356539
http://asmnt.sagepub.com
SAGE

Where Do We Draw the Line? Assigning Cases to Subsamples for MAMBAC, MAXCOV, and MAXEIG Taxometric Analyses

Glenn D. Walters[1] and John Ruscio[2]

Abstract

There are several important decisions that must be made when implementing taxometric procedures such as mean above minus below a cut (MAMBAC), maximum covariance (MAXCOV), and maximum eigenvalue (MAXEIG). A Monte Carlo study was performed with 10,000 (5,000 categorical, 5,000 dimensional) samples to examine 5 ways to locate the first and last MAMBAC cuts and 24 ways to perform MAXCOV and MAXEIG. For MAMBAC, there was little difference across conditions, with slightly more accurate results obtained when a small, fixed number of cases (n = 10 or 25) was located beyond the most extreme cuts. For MAXCOV and MAXEIG, the results were more palpable: MAXCOV slightly outperformed MAXEIG, windows achieved significantly better results than intervals, and a larger number of cases per subsample were associated with more accurate results. Alcohol misuse data obtained from a group of male prisoners were used to illustrate relationships observed in the Monte Carlo study.

Keywords

taxometrics, MAMBAC, MAXCOV, MAXEIG, cuts, intervals, windows

The taxometric method pioneered by Meehl (1995) and colleagues (e.g., Waller & Meehl, 1998) is designed to assess whether the latent structure of a construct is more appropriately modeled as categorical (two latent classes) or dimensional (one or more latent factors). The primary output of taxometric procedures is graphical, with different curve shapes expected for categorical and dimensional data, and the hallmark of this data-analytic approach is the examination of consistency across results from nonredundant analyses (Meehl, 1995). When judging the shapes of curves or assessing numerical results, the investigator's task is to determine whether the results are sufficiently consistent to support a structural conclusion or the results are ambiguous and judgment should be withheld (Meehl, 2004). Historically, taxometric results have been interpreted fairly subjectively, though recent developments such as the parallel analysis of categorical and dimensional comparison data and the calculation of a quantitative index of curve fit (the comparison curve fit index, or CCFI; Ruscio & Kaczetow, 2009; Ruscio, Ruscio, & Meron, 2007), which indicates the degree to which the empirical data curve fits the categorical and dimensional comparison data curves, have made the taxometric method more objective. Nonetheless, applying taxometric procedures to one's data is far from an automatic process.

A number of implementation decisions must be made to perform a taxometric analysis. Several of the more important implementation decisions include assigning the available variables to the required roles of input and output indicators, placing cuts along the input indicator or dividing cases into a series of ordered subsamples along the input indicator, and graphing the results (Ruscio, Haslam, & Ruscio, 2006). Walters and Ruscio (2009) recently examined the first of these decision points (i.e., arranging the variables into input and output indicators). Traditionally, a single indicator variable was placed along the x-axis to serve as the input indicator in a mean above minus below a cut (MAMBAC; Meehl & Yonce, 1994), maximum covariance (MAXCOV; Meehl & Yonce, 1996), or maximum eigenvalue (MAXEIG; Waller & Meehl, 1998) analysis. Beginning with Gangestad and Snyder (1985), many investigators have set aside the required output indicator(s) for analysis and summed the remaining variables to form a composite input indicator. Using both a large-scale Monte Carlo analysis

[1]Federal Correctional Institution–Schuylkill, Minersville, PA, USA
[2]The College of New Jersey, Ewing, NJ, USA

Corresponding Author:
Glenn D. Walters, Psychology Services, FCI–Schuylkill, PO Box 700, Minersville, PA 17954-0700, USA
Email: gwalters@bop.gov

and actual data from a structured diagnostic interview, Walters and Ruscio (2009) determined that summing variables to form a composite input indicator improved the accuracy of the MAMBAC procedure only slightly and actually had a significant detrimental effect on the accuracy of the MAXCOV and MAXEIG procedures. In the same study, indicators composed of four or more ordered categories achieved MAMBAC, MAXCOV, and MAXEIG results on par with continuous indicators.

The MAMBAC procedure sorts cases along an input indicator that forms the *x*-axis of a graph, with cuts placed at a series of evenly spaced locations to plot mean differences on the output indicator along the *y*-axis (Meehl & Yonce, 1994). Each mean difference is calculated as the mean for cases above the cut minus the mean for cases below the cut. Categorical data are expected to yield a peaked MAMBAC curve, whereas dimensional data are expected to yield a concave curve that often rises sharply at one or both ends (Meehl & Yonce, 1994). To implement MAMBAC, one must choose the number of cutting scores and the location of the most extreme cuts. The default values for Ruscio's (2009) taxometric program to perform MAMBAC analyses are to use 50 cuts and to reserve $n = 25$ cases at each extreme for the first and last cuts. Ruscio et al. (2006) contend that going beyond 50 cuts will impose a computational burden that is unlikely to be reciprocated by a noticeable increase in accuracy, although this is an empirical question that we do not address in the current study. The MAMBAC implementation decision investigated in this study is how many cases should be reserved at the extremes to produce the most informative results.

The MAXCOV procedure divides cases into a series of ordered subsamples along the input indicator, which forms the *x*-axis of a graph. Within each subsample, the covariance between two output indicators is calculated, and these values are plotted such that covariance forms the *y*-axis of the graph (Meehl & Yonce, 1996). The MAXEIG procedure is a multivariate generalization of the MAXCOV procedure (Waller & Meehl, 1998). Rather than calculating the covariance of two output indicators within each subsample, one calculates the first (largest) eigenvalue of the covariance matrix of two or more output indicators, where the covariance matrix is the usual variance–covariance matrix with the variances replaced by zeros. For both procedures, categorical data are expected to yield a peaked curve, whereas dimensional data normally fail to show evidence of a discernable peak (Meehl & Yonce, 1996; Waller & Meehl, 1998).

To implement these procedures, several decisions must be made. First, should one calculate covariances among pairs of output indicators (and thereby perform MAXCOV) or calculate eigenvalues among two or more output indicators (and thereby perform MAXEIG)? When there are more than three variables available to serve as indicators, MAXEIG will include more variables in each analysis than MAXCOV, with the potential benefit of yielding more powerful results. However, eigenvalues calculated from more than two indicators are subject to greater sampling error than covariances calculated from two indicators, all else being equal. We performed both MAXCOV and MAXEIG analyses of the same target data sets to determine whether there is a net benefit to using either procedure.

Second, should one divide cases into subsamples along the input indicator using nonoverlapping intervals or overlapping windows? For example, consider a sample with $N = 1{,}000$ cases divided into 10 intervals of $n = 100$. These would contain cases 1 to 100, 101 to 200, 201 to 300, . . . , 901 to 1,000. Dividing this same sample into windows with $n = 100$ that overlap 90% with their neighbors affords 91 windows (see Waller & Meehl, 1998, for formulas relating sample size, number of windows, amount of overlap, and subsample size). These would contain cases 1 to 100, 11 to 110, 21 to 120, . . . , 901 to 1,000. This series of windows includes the original 10 decile intervals, along with 9 additional windows between each, for a total of 91 windows (Ruscio et al., 2006). We expect that using windows will be preferable to using intervals because one can obtain more data points and better flesh out the shape of a MAXCOV or MAXEIG curve with no increase in the amount of sampling error within each subsample.

Third, should one choose the number of subsamples (intervals or windows), which then determines the number of cases within each one, or choose the number of cases within each subsample, which then determines the number of subsamples? In other words, as a general rule, is it wiser to set a fixed number of subsamples or a fixed number of cases per subsample? For example, one might choose to use 25 intervals, in which case a sample of $N = 1{,}000$ cases would provide $n = 40$ cases within each interval. Alternatively, one might choose to use intervals with $n = 50$ cases, in which case this would provide 20 intervals. The same type of relationship holds for windows: One might choose to use 50 windows with $n = 169$, or one might choose to use $n = 50$ and thereby obtain 191 windows. We expect choosing the number of subsamples to be a better approach, provided that one chooses a sufficient but not excessive number of data points to establish the shape of a curve.

Finally, how many subsamples (or cases) should be implemented when doing MAXCOV or MAXEIG? Currently, the number of subsamples (or cases) in a MAXCOV or MAXEIG analysis is a matter of personal preference. The default values in Ruscio's (2009) taxometric program that performs MAXCOV and MAXEIG analyses are 15 intervals and 50 windows, but these are based on educated guesses rather than data.

In an effort to establish empirically based default values that can guide taxometric investigations, we studied the

implementation decisions outlined above in MAMBAC, MAXCOV, and MAXEIG analyses of large numbers of categorical and dimensional target data sets. Specifically, this study was designed to investigate the effect of adjusting the location of the most extreme cuts in MAMBAC analyses and the type and number of subsamples in MAXCOV and MAXEIG analyses. Parallel analyses of comparison data were performed so that the CCFI could be calculated for each variant of each procedure. We compared implementation methods by examining their accuracies using several criteria, including receiver operating characteristic (ROC) curves and accuracy scores based on the application of single or dual thresholds for CCFI values. Applying a single threshold allows an evaluation of the percentage of samples whose structure was identified correctly, and applying dual thresholds allows intermediate CCFI values to be scored as ambiguous (partial credit) rather than correct (full credit) or incorrect (no credit). Hence, the ROC analyses were used to assess the overlap between CCFI scores for categorical and dimensional samples and the single- and dual-threshold criteria were used to assess the practical utility of the CCFI in reaching conclusions using prespecified thresholds.

Study 1

Method

Design and data generation. Using an iterative technique developed by Ruscio and Kaczetow (2008), we created 10,000 target data sets for a Monte Carlo study by crossing several important data parameters. Latent structure (categorical or dimensional) was the only parameter that was systematically varied across samples, with 5,000 samples for each structure. All other data parameters were sampled randomly and independently from uniform distributions.

To construct categorical data, the following data parameters (and ranges of values) were varied: sample size ($N = 300$ to 1,000), number of indicators (3 to 8), taxon[1] base rate ($P = .10$ to .50), indicator validity ($d = 1.25$ to 2.00), within-group correlations ($r = .00$ to .30), asymmetry ($g = .00$ to .30), tail weight ($k = .00$ to .15), and taxon:complement variance ratio ($VR = 1$ to 4). Categorical data sets were generated with Ruscio and Kaczetow's (2008) iterative technique, with N cases sampled from a g-and-h distribution (Hoaglin, 1985) with $\mu = 0$ and $\sigma = 1$ and a correlation matrix in which all indicators correlated r with one another in the taxon and complement groups. A proportion (P) of cases were then randomly selected and classified as taxon members, with the remaining cases serving as the complement. Finally, scores in the taxon group were multiplied by the VR and the classes separated by adding a constant to the taxon group that achieved the desired group separation d.

Because several of the parameters used to create categorical data do not correspond to parameters in the dimensional model (i.e., P, d, r, and VR), these parameters were replaced by indicator correlations sampled from a uniform distribution (r_{xy}) that ranged from .15 to .65. The iterative algorithm of Ruscio and Kaczetow (2008) was used to sample N cases from a g-and-h distribution with $\mu = 0$ and $\sigma = 1$, and a correlation matrix in which all of the indicators correlated r_{xy} with one another. Subsequent analysis and review revealed that the Ruscio and Kaczetow (2008) algorithm generated categorical and dimensional data sets with the intended indicator correlations, distributions, and variance ratios.

Taxometric analyses. MAMBAC, MAXCOV, and MAXEIG were each performed as described earlier, with one variable serving as the input indicator and either one (MAMBAC), two (MAXCOV), or the remaining (MAXEIG) indicator variables serving as the output indicator(s). For MAMBAC analyses, 50 cuts were located, with the number of cases reserved beyond the most extreme cuts varied across five conditions. The number of cases was either independent of N and fixed at $n = 10$, 25, or 50 cases or calculated as 5% or 10% of N. For each target data set, MAMBAC was performed five times, and for each of these analyses the full panel of curves was averaged and accompanied by results for parallel analyses of categorical and dimensional comparison data (Ruscio et al., 2007; Ruscio & Kaczetow, 2009). This enabled the calculation of the CCFI, which quantified the relative fit of categorical and dimensional structural models. CCFI values >.50 are indicative of categorical structure whereas CCFI values <.50 are indicative of dimensional structure. Previous Monte Carlo research strongly supports the utility of the CCFI in correctly identifying the structure of data generated using either of these models (Ruscio et al., 2007; Ruscio & Kaczetow, 2009; Ruscio & Marcus, 2007; Ruscio, Walters, Marcus, & Kaczetow, in press; Walters & Ruscio, 2009). In the present study, five CCFI values were obtained for MAMBAC analyses of each target data set.

For MAXCOV and MAXEIG analyses, four factors were varied across a total of 24 conditions: (a) calculating covariances between pairs of output indicators (MAXCOV) versus calculating eigenvalues between all available output indicators (MAXEIG), (b) dividing cases into subsamples using intervals versus windows, (c) fixing the number of subsamples versus fixing the number of cases per subsample, and (d) using a small, medium, or large number of subsamples or cases. Specifically, the following combinations of factors (b) through (d) were studied for MAXCOV and MAXEIG analyses: 10, 25, or 40 intervals; intervals with $n = 10$, 25, or 50; 25, 50, or 100 windows; windows with $n = 10$, 25, or 50. As in the MAMBAC analyses, the full panel of curves for each MAXCOV and MAXEIG analysis was averaged, parallel analyses of comparison data

were performed, and the CCFI was calculated. This resulted in 24 CCFI values for each target data set, 12 for MAXCOV analyses and 12 for MAXEIG analyses.

Outcome measures. The ability of the CCFI score to differentiate between categorical and dimensional target data sets was evaluated using ROC analyses in which area under the curve (AUC) values were calculated. This provides a measure of accuracy that is independent of threshold. In addition, accuracy scores were calculated by applying a single threshold or dual thresholds to CCFI values. When a single threshold was applied, CCFI < .50 indicated dimensional structure, CCFI > .50 indicated categorical structure, and accuracy was scored as 1 for correct and 0 for incorrect structural identification. When dual thresholds were applied, intermediate CCFI values were scored as ambiguous (.5); this was done using narrow dual thresholds (CCFIs of .45 and .55) and broad dual thresholds (CCFIs of .40 and .60). Scores on each threshold-dependent accuracy measure were entered into a repeated measures analysis of variance (ANOVA) to test whether results differed across conditions for each taxometric procedure. For MAMBAC, there was a single within-subjects factor (location of the most extreme cuts), and for MAXCOV and MAXEIG there were four within-subjects factors (MAXCOV vs. MAXEIG, intervals vs. windows, fixed number of subsamples vs. fixed number of cases, and small, medium, or large number of cases per subsample). Because the use of our dichotomous (single threshold) and trichotomous (dual thresholds) outcome measures violated the normality assumption and our large sample size could have resulted in significant effects of trivial magnitude, we focused on the estimated effect size (using partial η^2, calculated using SPSS) for each main effect or interaction rather than its statistical significance. The usual parametric assumptions are not required to calculate and evaluate this measure of effect size.

Results

MAMBAC. The AUC and classification results for the 5 MAMBAC, 12 MAXCOV, and 12 MAXEIG conditions examined in this study are listed in Table 1. For MAMBAC, accuracy was very high for all conditions (each AUC ≥ .997). For both single-threshold and dual-thresholds percentage correct (Table 1) and accuracy scores (Figure 1, top graph), there was a slight advantage for a small, fixed number of cases (n = 10) reserved beyond the extreme cuts. For this condition, MAMBAC identified structure correctly 96.5% of the time using a single threshold, 98.5% of the time using narrow dual thresholds (after setting aside 5.7% of the results as ambiguous), and 99.5% of the time using broad dual thresholds (after setting aside 13.1% of the results as ambiguous). A willingness to reserve judgment for ambiguous results would be rewarded with an increase in accuracy among those that remain.

Repeated measures ANOVAs were performed for MAMBAC, once for each of the three accuracy scores (single threshold, narrow dual thresholds, and broad dual thresholds). Consistent with the results presented above, the five MAMBAC conditions differed only slightly from one another, with trivial effects for all three accuracy scores (each partial η^2 < .01; see Table 2). As a follow-up, mean accuracy scores were plotted for the 5,000 categorical data sets broken down by four levels of taxon base rates (.10 to .20, .20 to .30, .30 to .40, and .40 to .50); these results are shown in the bottom three graphs of Figure 1. Accuracy tended to be higher for lower base rates, and the results continued to differ very little across conditions. Taken as a whole, these results provide little reason to prefer any particular location of the most extreme cuts in MAMBAC analyses. By a very small margin, overall accuracy was highest when n = 10 cases were reserved beyond the first and last cuts.

MAXCOV and MAXEIG. For MAXCOV and MAXEIG, accuracy was a bit lower than for MAMBAC but still quite high (each AUC ≥. 975; see Table 1). Results diverged most sharply when dual thresholds were applied (see Figure 2, graphs on the left). Regardless of threshold(s), however, accuracy was greatest using MAXEIG analyses with 25 windows; MAXCOV analyses with 25 windows yielded nearly identical results. For this condition, MAXEIG identified structure correctly 95.2% of the time using a single threshold, 97.3% of the time using narrow dual thresholds (after setting aside 5.8% of the results as ambiguous), and 98.3% of the time using broad dual thresholds (after setting aside 12.3% of the results as ambiguous). Here, too, reserving judgment for ambiguous results increased accuracy among the remainder of the data.

All four within-subjects factors were included in the repeated measures ANOVAs for MAXCOV/MAXEIG. Using the single-threshold accuracy score, all main effects and interactions were trivial in magnitude (each partial η^2 < .01; see Table 2). However, analyses using the dual-thresholds accuracy scores revealed nontrivial main effects for each of the four factors, such that accuracy was better for MAXCOV than for MAXEIG, for windows rather than for intervals, for a specified number of subsamples than for a specified number of cases per subsample, and for larger rather than for smaller numbers of cases per subsample. We included all possible interaction effects in the ANOVAs. Approximately half of these were trivial in magnitude (partial η^2 < .01), and even among those that were not (.01 < partial η^2 < .35), all effects were monotonic in that they did not qualify the direction of any main effects. As noted above, an inspection of results suggests that the most accurate results were obtained using 25 windows, with very little difference between MAXCOV and MAXEIG under that condition. Perhaps surprisingly, these findings were not qualified by an interaction with taxon base rates,

Table 1. CCFI Accuracy Results for ROC and Classification Analyses Across Implementation Conditions

Procedure and Implementation	AUC	Single Threshold	Narrow Dual Thresholds	Broad Dual Thresholds
MAMBAC-10	.998 (.998-.999)	96.5	98.5 (5.7)	99.5 (13.1)
MAMBAC-25	.998 (.998-.999)	96.2	98.1 (6.1)	99.3 (13.5)
MAMBAC-50	.997 (.997-.998)	95.6	97.7 (6.3)	98.8 (13.2)
MAMBAC-5P	.998 (.998-.999)	96.0	98.3 (6.6)	99.2 (13.8)
MAMBAC-10P	.997 (.996-.998)	95.1	97.5 (6.6)	98.9 (14.4)
MAXCOV-10i	.984 (.982-.986)	94.0	96.6 (7.8)	98.1 (17.7)
MAXCOV-25i	.979 (.977-.982)	93.1	97.3 (14.7)	99.0 (34.5)
MAXCOV-40i	.980 (.978-.982)	93.0	98.3 (22.6)	99.7 (51.9)
MAXCOV-(n10)i	.978 (.976-.981)	93.0	98.8 (33.0)	99.8 (72.9)
MAXCOV-(n25)i	.979 (.977-.982)	93.4	97.1 (13.2)	98.9 (32.9)
MAXCOV-(n50)i	.984 (.982-.986)	93.8	96.6 (8.5)	98.3 (19.2)
MAXEIG-10i	.982 (.980-.984)	93.9	96.6 (8.6)	98.0 (18.7)
MAXEIG-25i	.979 (.977-.981)	92.9	97.6 (17.3)	99.3 (39.1)
MAXEIG-40i	.977 (.975-.980)	92.3	98.6 (28.9)	99.8 (58.5)
MAXEIG-(n10)i	.975 (.972-.977)	91.6	98.9 (44.8)	99.9 (79.3)
MAXEIG-(n25)i	.979 (.977-.982)	93.0	97.5 (16.3)	99.1 (38.1)
MAXEIG-(n50)i	.982 (.980-.985)	93.7	96.6 (9.1)	98.4 (20.7)
MAXCOV-25w	.988 (.987-.990)	95.0	97.1 (5.6)	98.2 (12.4)
MAXCOV-50w	.986 (.984-.988)	94.3	96.8 (7.1)	98.3 (16.3)
MAXCOV-100w	.984 (.982-.986)	94.0	96.9 (9.2)	98.8 (22.9)
MAXCOV-(n10)w	.984 (.982-.986)	93.8	99.4 (36.4)	99.9 (77.1)
MAXCOV-(n25)w	.983 (.981-.985)	94.0	97.8 (14.8)	99.4 (39.0)
MAXCOV-(n50)w	.984 (.983-.986)	94.2	97.0 (9.4)	98.8 (23.5)
MAXEIG-25w	.989 (.988-.991)	95.2	97.3 (5.8)	98.3 (12.3)
MAXEIG-50w	.986 (.984-.988)	94.2	96.8 (7.2)	98.4 (16.3)
MAXEIG-100w	.984 (.982-.986)	93.9	97.1 (10.4)	98.8 (24.1)
MAXEIG-(n10)w	.980 (.978-.982)	92.1	99.6 (49.4)	99.9 (82.5)
MAXEIG-(n25)w	.982 (.980-.984)	93.6	98.2 (18.4)	99.5 (44.0)
MAXEIG-(n50)w	.984 (.982-.986)	93.9	97.0 (10.1)	99.0 (24.7)

Note: ROC = receiver operating characteristic; MAMBAC = mean above minus below a cut (10 to 50 cases in extreme cuts or 5% to 10% of cases in extreme cuts); MAXCOV = maximum covariance; MAXEIG = maximum eigenvalue; i = intervals; w = windows; n10 = 10 cases per subsample; n25 = 25 cases per subsample; n50 = 50 cases per subsample; AUC = area under the ROC curve (calculated using distribution-free geometric formula) with asymptotic 95% confidence intervals in parentheses; number in single and dual thresholds columns is percentage of correctly classified cases; in dual threshold columns, percentage of indeterminate cases is indicated in parentheses; single threshold was placed at comparison curve fit index (CCFI) = .50, narrow dual thresholds were placed at CCFIs of .45 and .55, and broad dual thresholds were placed at CCFIs of .40 and .60.

although none of the base rates examined in this study were below .10. The right side of Figure 2 presents MAXCOV results for analyses of the 5,000 categorical data sets across four levels of base rates. For each accuracy score and within each range of base rates, using 25 windows yielded the most accurate results.

Discussion

From the results of this Monte Carlo analysis we conclude that MAMBAC performs as well with a small, fixed number of cases placed at the extremes as when a large, fixed number of cases or a portion of cases are placed at the extremes; on the other hand, a MAXCOV analysis yields slightly more accurate results than a MAXEIG analysis, overlapping windows are more useful than nonoverlapping intervals, and the best results are obtained when the number of subgroups is low and the number of cases per subgroup is high. However, because the indicators included in this Monte Carlo analysis were continuous we do not know how well the results generalize to analyses using indicators composed of a small number of ordered categories. To illustrate the findings and examine their application with ordered categorical data, a study was performed using real data from a previously published taxometric study of alcohol misuse disorders (Walters, 2008). We emphasize that these real data analyses are provided for illustrative purposes rather than for confirmation or generalization of the Monte Carlo results from Study 1.

Study 2

Method

Participants. The sample for this study comprised 1,193 male federal prison inmates who were administered a

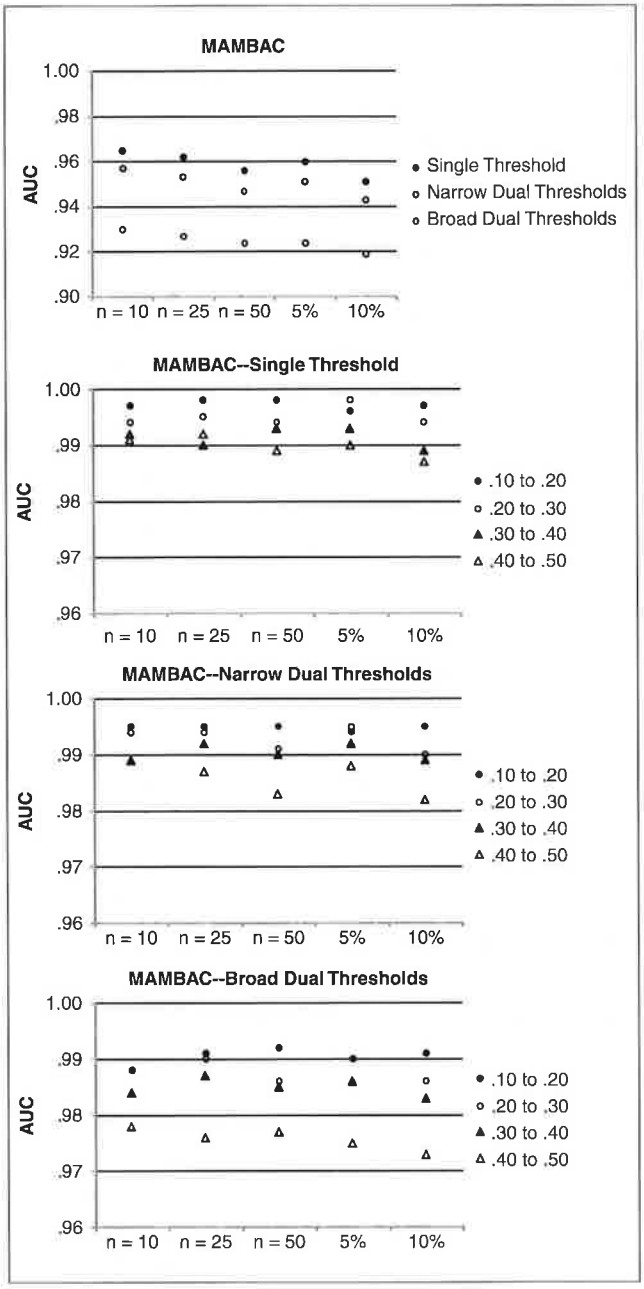

Figure 1. AUC results by number and percentage of cases beyond extreme cuts in MAMBAC

Note: AUC = area under receiver operating characteristic curve. Top graph shows results for all 10,000 samples of target data and bottom three graphs show results for all 5,000 samples of categorical target data broken down by the base rate of the smaller group. The single threshold was placed at comparison curve fit index (CCFI) = .50, narrow dual thresholds were placed at CCFIs of .45 and .55, and broad dual thresholds were placed at CCFIs of .40 and .60. For the single threshold, scoring was correct = 1 and incorrect = 0; for narrow or broad dual thresholds, scoring was correct = 1, ambiguous = .5, incorrect = 0.

structured interview designed to determine their eligibility for a comprehensive drug treatment program. Participants averaged 35.35 years of age (SD = 9.68), had 11.85 years of education (SD = 1.97), and three quarters of the sample (66.2%) satisfied *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (*DSM-IV*; American Psychiatric Association, 1994) criteria for alcohol dependence. The ethnic breakdown was 40.2% Caucasian American, 38.5% African-American, 16.2% Hispanic, 3.7% Asian, and 1.5% Native American, and the marital breakdown was 57.8% single, 28.8% married, 12.7% divorced/separated, and 0.7% widowed.

Measure. Indicators were derived from a structured interview composed of 15 dichotomous items that assess the seven *DSM-IV* criteria for alcohol dependence and four *DSM-IV* criteria for alcohol abuse. This structured interview was developed by psychologists in the Federal Bureau of Prisons to determine inmate eligibility for a comprehensive drug treatment program. There was both a theoretical (content analysis) and empirical (factor analysis) rationale for organizing the 15 items into three indicators spanning 5 to 7 ordered categories each (see Walters, 2008). Indicator 1 comprised four dichotomous items measuring two *DSM-IV* criteria for alcohol dependence (tolerance; withdrawal) used to form a 5-point scale (range = 0-4, M = 1.78, SD = 1.54) with adequate internal consistency (α = .82). Indicator 2 comprised dichotomous items measuring three *DSM-IV* alcohol-dependence criteria (larger amounts and periods of use; unsuccessful attempts to cut down; time spent obtaining, using, and recovering from effects) and one *DSM-IV* alcohol-abuse criterion (legal problems) used to form a 7-point scale (range = 0-6, M = 3.13, SD = 2.34) with good internal consistency (α =.87). Indicator 3 comprised dichotomous items measuring two *DSM-IV* alcohol-dependence criteria (reduction in social, occupational, or recreational activities; continued use despite physical or psychological problems) and three *DSM-IV* alcohol-abuse criteria (failure to meet role obligations, physically hazardous activities, social/interpersonal problems) used to form a 6-point scale (range = 0-5, M = 2.91, SD = 2.08) with good internal consistency (α = .90). Inter-rater reliability was found to be adequate for 40 randomly selected cases independently interviewed by a second interviewer 2 to 8 weeks after the original interview (Intraclass Correlation Coefficient (ICC) = .60, $p < .001$).

Procedure. Taxometric analyses were performed using Ruscio's (2009) taxometric programs. MAMBAC was performed using 50 cuts with n = 10, 25, 50, 60 (5% of the total N) or 119 (10% of the total N) cases reserved beyond the first and last cuts. MAXCOV was performed rather than MAXEIG because it achieved slightly more accurate results in an earlier Monte Carlo study (Ruscio et al., in press) as well as in Study 1 of the current investigation. MAXCOV was performed with 25, 50, and 100 windows at 90% overlap and then with 10, 25, and 40 nonoverlapping intervals. For all analyses, 10 internal replications were performed to minimize the obfuscating effect of tied scores (Ruscio et al., 2006).

Table 2. Effect Size (Partial η^2) for ANOVAs Across Design Factors

	Single Threshold	Narrow Dual Thresholds	Broad Dual Thresholds	Pattern of Results
MAMBAC	.001	.002	.001	
MAXCOV and MAXEIG				
A	.003	.075	.074	MAXCOV > MAXEIG
B	.005	.234	.503	subsamples > cases
C	.006	.032	.134	windows > intervals
D	.009	.278	.550	large *n* > small *n*
A × B	.002	.036	.010	
A × C	.000	.003	.012	
A × D	.002	.048	.015	
B × C	.003	.111	.349	
B × D	.001	.166	.274	
C × D	.000	.006	.069	
A × B × C	.000	.008	.008	
A × B × D	.001	.020	.001	
A × C × D	.000	.001	.001	
B × C × D	.000	.028	.077	
A × B × C × D	.000	.002	.001	

Note: ANOVA = analysis of variance; MAMBAC = mean above minus below a cut; MAXCOV = maximum covariance; MAXEIG = maximum eigenvalue. Three series of repeated measures ANOVAs were run using different dependent variables. The single threshold was placed at comparison curve fit index (CCFI) = .50, narrow dual thresholds were placed at CCFIs of .45 and .55, and broad dual thresholds were placed at CCFIs of .40 and .60. For the single threshold, scoring was correct = 1 and incorrect = 0; for narrow or broad dual thresholds, scoring was correct = 1, ambiguous = .5, incorrect = 0. Factors included in the MAXCOV and MAXEIG ANOVAs are coded as follows: A = MAXCOV versus MAXEIG; B = specified number of subsamples versus specified number of cases per subsample; C = windows versus intervals; D = large, medium, or small number of cases per subsample. Main effects are briefly described, and all interactions were monotonic (none qualified the main effects).

Results

To estimate data parameters, cases were classified according to whether or not they met *DSM-IV* criteria for alcohol dependence. Across the 12 conditions, estimated indicator validity was comparatively strong ($d = 2.37$ to 3.38, $M = 3.02$) and estimated within-group indicator correlations were substantial (for taxon, mean $r = .28$ and for complement, mean $r = .35$) but considerably smaller than full-sample correlations (mean $r = .75$). For MAMBAC, the CCFI results were all very high (.872 to .912), providing strong support for an inference of categorical structure regardless of how many cases were reserved beyond the first and last cuts. Figure 3 shows the graphical results, in the context of those for categorical and dimensional comparison data. The most striking feature of these graphs is the similarity of the results. Reserving different numbers of cases beyond the first and last cuts has very little effect on most of the 50 data points that count equally in the calculation of the CCFI, which demonstrates why this implementation decision was shown not to be very important in Study 1.

For MAXCOV, on the other hand, CCFIs spanned a wider range (.665 to .871), underscoring the importance of implementing the procedure most effectively. As shown in Figure 4, CCFIs were larger for windows (graphs in left column) than for intervals (graphs in right column) and for fewer subsamples (graphs in top row) than for more subsamples (graphs in bottom row). By far, the strongest results were obtained using the smallest number of windows (25) and the weakest results were obtained using the largest number of intervals (40). We performed MAXCOV using even smaller windows and intervals (allowing *n* to drop to 25 and then to 10), and CCFI values continued to decline. Here, as with the MAMBAC analyses, results parallel those obtained in the Monte Carlo analyses reported in Study 1 and suggest that these findings may extend to ordered categorical data.

Discussion

If the latent structure of alcohol use disorders, is, in fact, categorical—and there is both an adequate rationale (Lenzenweger, 2004) and empirical support in six out of the seven taxometric studies on alcohol-use disorders (Dana, 1990; Green, Ahmed, Marcus, & Walters, 2009; Walters, 2008, 2009; Walters, Diamond, & Magaletta, in press; Walters, Henning, Negola, & Fricke, 2009; see Slade, Grove, & Teeson, 2009 for an exception) for this assumption—then the current results support the relationships witnessed in Study 1 whereby there is little difference across MAMBAC implementations but substantial differences across MAXCOV implementations. For the latter, windows produced stronger results than intervals and a smaller number of subgroups (with a comparatively large *n* in each) yielded stronger results than a larger number of subgroups (with a comparatively small *n* in each).

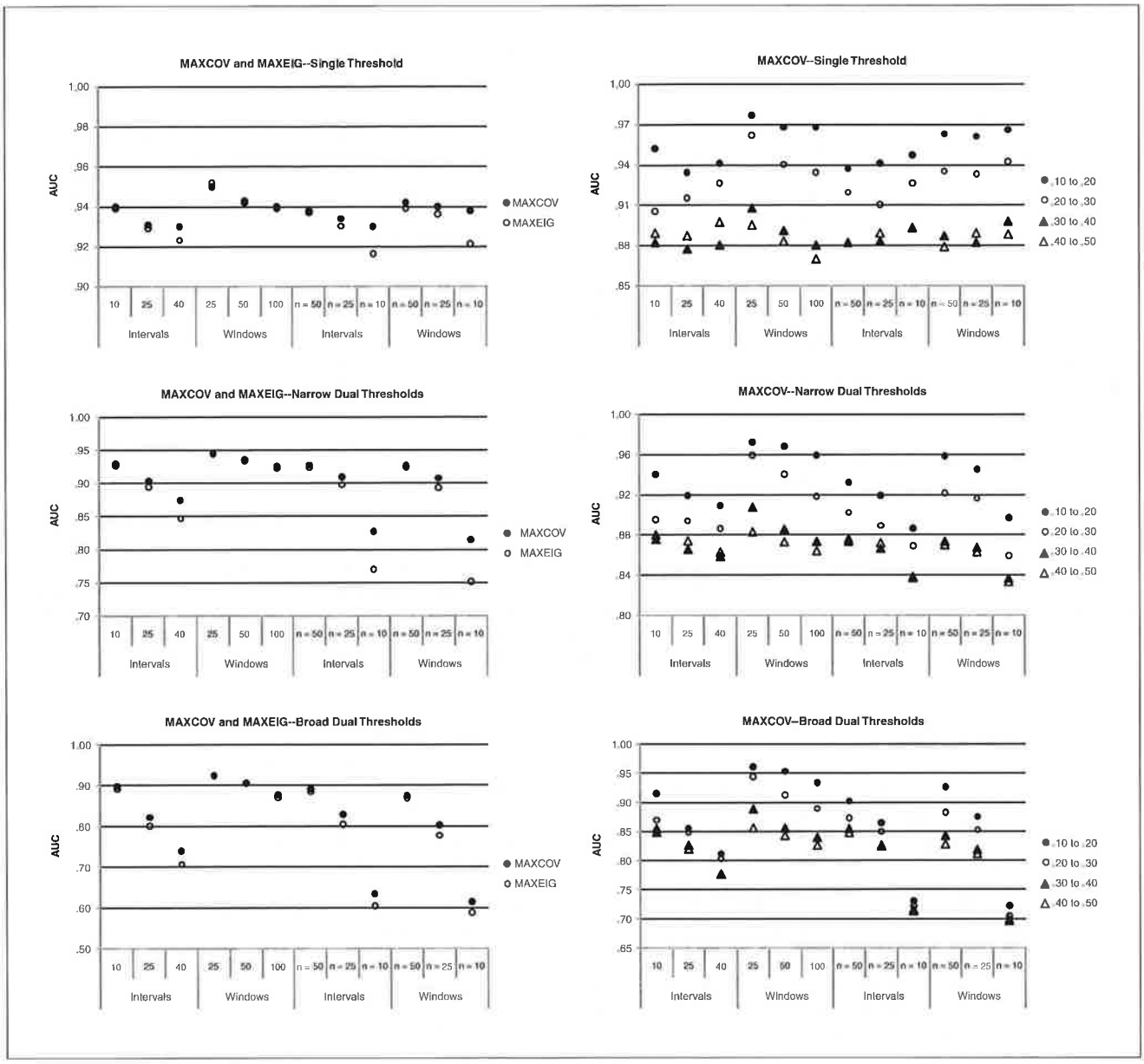

Figure 2. AUC results by type and number of subsamples in MAXCOV and MAXEIG

Note: AUC = area under receiver operating characteristic curve. Three graphs at left show results for all 10,000 samples of target data, three graphs at right show results for all 5,000 samples of categorical target data broken down by the base rate of the smaller group. The single threshold was placed at Comparison Curve Fit Index (CCFI) = .50, narrow dual thresholds were placed at CCFIs of .45 and .55, and broad dual thresholds were placed at CCFIs of .40 and .60. For the single threshold, scoring was correct = 1 and incorrect = 0; for narrow or broad dual thresholds, scoring was correct = 1, ambiguous = .5, incorrect = 0.

General Discussion

The results of the Monte Carlo study (Study 1) indicate that it matters only slightly how many cases one reserves for the extreme cuts of a MAMBAC analysis but that the number and type of subsamples one employs in a MAXCOV or MAXEIG analysis are more consequential. The impact of number and type of subsamples in MAXCOV was illustrated in a study using actual data from participants administered a structured interview based on *DSM-IV* alcohol dependence and abuse criteria (Study 2). MAMBAC seems to achieve slightly more accurate results when a small, fixed number of cases are placed at the extremes rather than when a large, fixed number of cases or a portion of cases are placed at the extremes. Restricting analyses to categorical

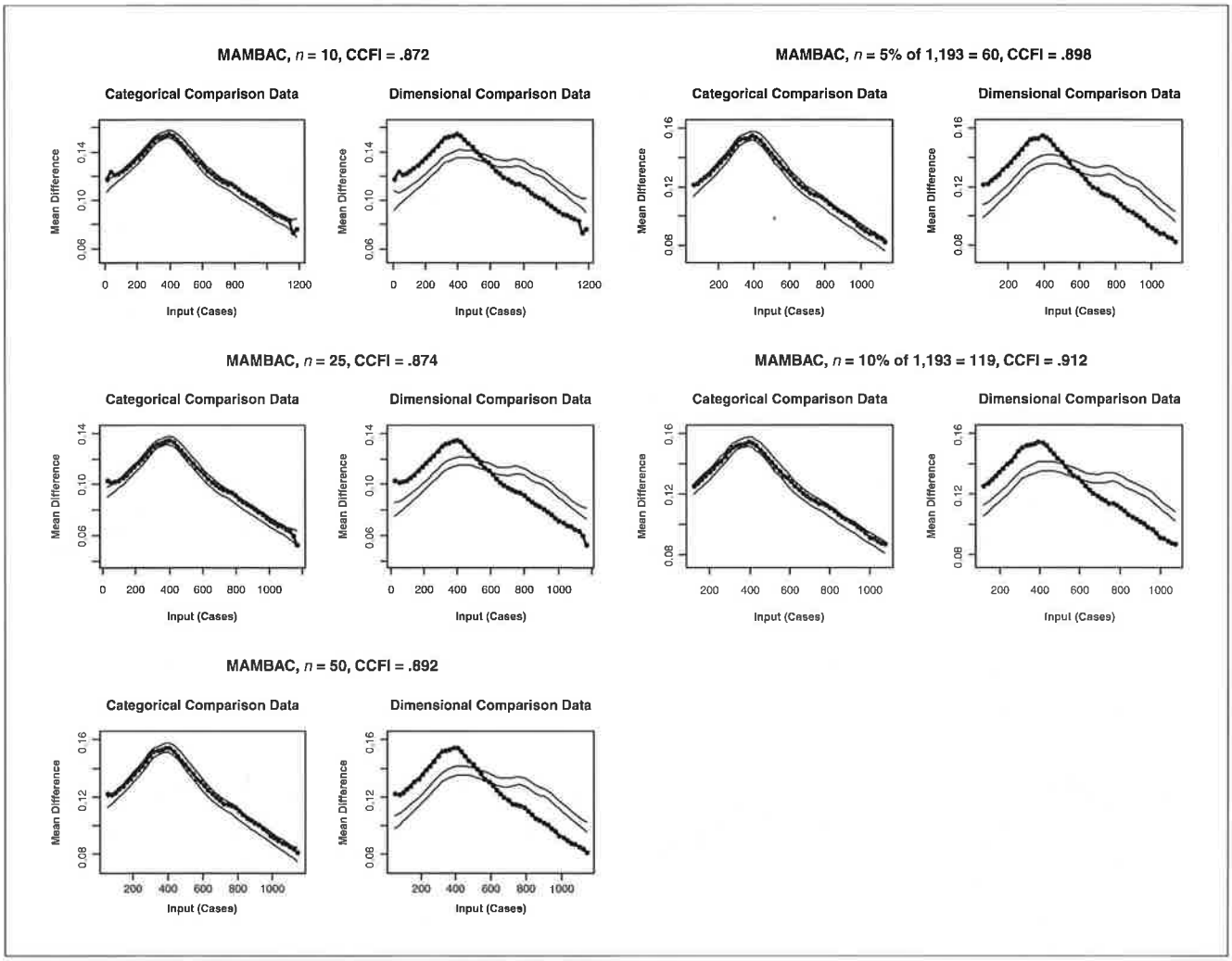

Figure 3. Results for MAMBAC analyses in Study 2, with differing numbers of cases reserved beyond the first and last cuts
Note: CCFI = comparison curve fit index.

data sorted into subsets according to taxon base rates did nothing to alter this conclusion.

In repeated measures ANOVAs examining differences in accuracy across implementations, several of the effect sizes for MAXCOV/MAXEIG were much larger than those obtained with MAMBAC. Although there was a slight advantage for MAXCOV over MAXEIG, there was a much larger advantage for windows over intervals and even stronger effects for fewer number of subgroups and larger numbers of cases per subgroup. Considering these findings as a whole, they suggest that dividing cases into a relatively small number of windows can play an important role in promoting the accuracy of MAXCOV or MAXEIG. This conclusion, too, held across taxon base rates in follow-up analyses of categorical data sets.

The present findings parallel those observed in Walters and Ruscio's (2009) study: MAMBAC appears to be less sensitive to variations in its implementation than MAXCOV/MAXEIG. This may be a consequence of how results for each procedure are calculated. Each cut in a MAMBAC analysis includes the entire sample but each subsample in a MAXCOV/MAXEIG analysis includes only a portion of the overall sample. Thus, there is greater opportunity for procedural variants to influence the results in MAXCOV/MAXEIG analyses than in MAMBAC analyses. Moreover, there are more implementation decisions to be made in MAXCOV/MAXEIG analyses than in MAMBAC analyses, hence more opportunities for choices to matter.

It should be noted that despite substantial effect sizes in the repeated measures ANOVAs for MAXCOV/MAXEIG, accuracy did not differ much between conditions. Because we used partial η^2 as the measure of effect size and the design included only repeated measures factors, nontrivial effect size estimates do not necessarily correspond to

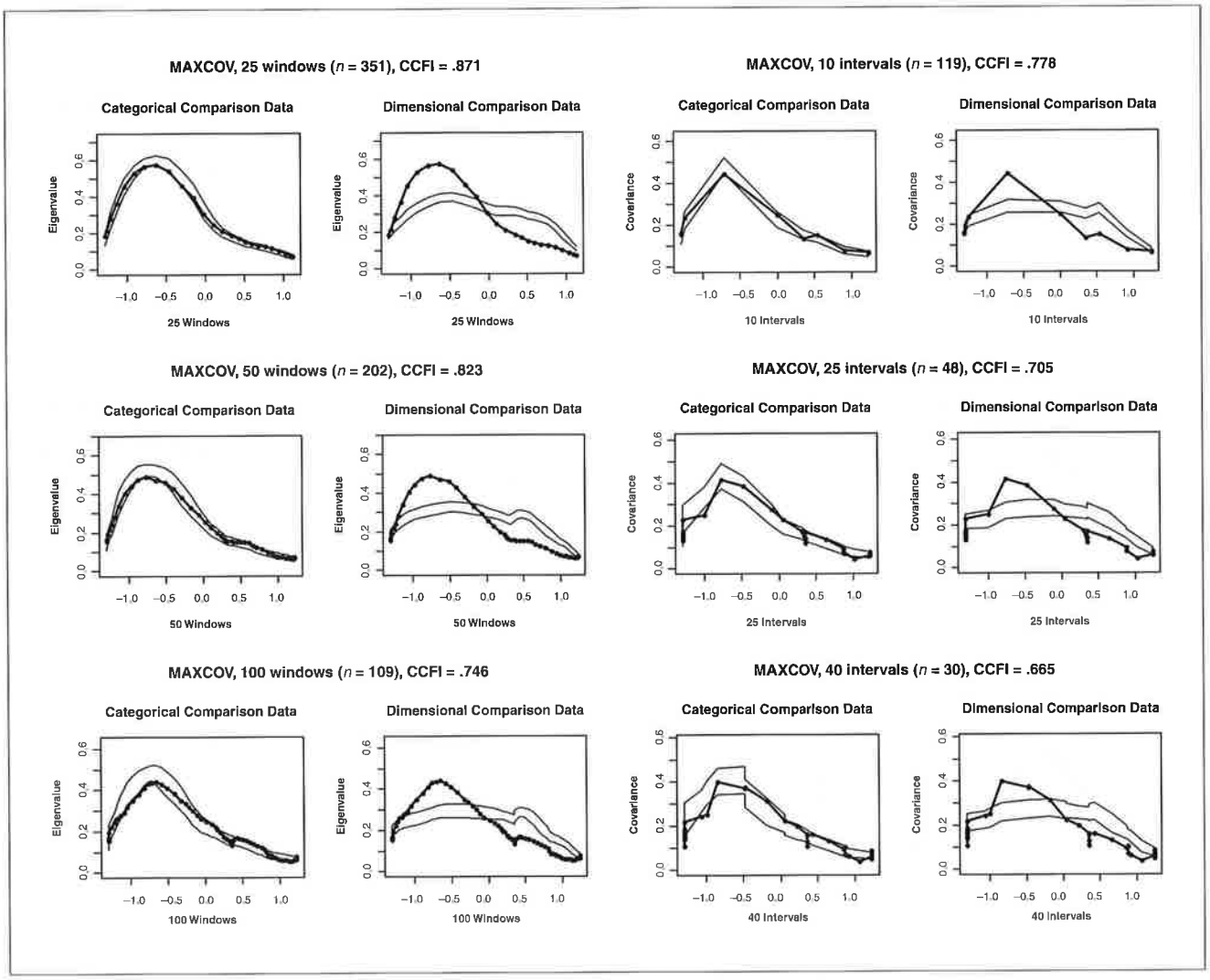

Figure 4. Results for MAXCOV analyses in Study 2, with varying numbers of windows (left) and intervals (right)
Note: CCFI = comparison curve fit index.

practically significant differences in the differentiation of categorical and dimensional data. AUC values ranged from .975 (for MAXEIG with intervals containing $n = 10$ cases each) to .989 (for MAXEIG with 25 windows), so discrimination was quite high for all 24 implementations of MAXCOV and MAXEIG. Because the 95% confidence intervals for AUC values never spanned a range wider than .005 units, even fairly small differences between conditions may appear to be statistically or practically significant.

The primary differences across conditions involve the percentage of CCFI values falling in the ambiguous range between the dual thresholds. When subsamples were smaller—either by selecting a smaller n per subsample or selecting a larger number of subsamples—the percentage of ambiguous results increased substantially. This increase in ambiguity, which would require withholding judgment more often, was offset by an increase in accuracy among the remaining data. However, our reading of these results suggests that the increase in accuracy was seldom large enough to represent an acceptable trade-off for the increase in ambiguity. For example, MAXEIG analyses with 25 windows achieved 98.3% accuracy at the cost of setting aside 12.3% of results as ambiguous when broad dual thresholds were applied. Applying the same thresholds to MAXEIG analyses with 100 windows increased accuracy by 0.5% (from 98.3% to 98.8%), but it nearly doubled the percentage of ambiguous results (from 12.3% to 24.1%). Investigators can choose the balance between accuracy and ambiguity with which they feel most comfortable, but we see little or no justification for using large numbers of subsamples or small-n subsamples. Of course, it could be argued that it really does not matter that small-n subsamples produce a relatively large

number of indeterminate results as long as they fall on the correct side of .50. The problem with this argument is that because CCFI results in the ambiguous range are significantly less accurate than CCFI results outside the ambiguous range,[2] an indeterminate category is required to avoid accepting results close to .50 as legitimate.

The present findings, especially the fact that using a small number of windows worked best even for the lowest taxon base rates in this study, has implications for the inchworm consistency test (Waller & Meehl, 1998). This test involves increasing the number of windows across a series of analyses to determine whether a cusp toward the right end of a MAXCOV or MAXEIG curve represents a small taxon or is an artifact of positively skewed indicators (Ruscio, Ruscio, & Keane, 2004). For categorical data with a low taxon base rate, a cusp may be transformed into a fully defined peak. For dimensional data, a cusp that stems from the influence of positive skew should remain even as the number of windows increases. The present results suggest that the CCFI can distinguish these structural possibilities at least as effectively with a small number of windows as with a large number of windows. This is consistent with the results of Ruscio and Marcus (2007), who found that the CCFI easily identified the categorical structure of data sets with base rates as low as $P = .05$ when reanalyzed using 50 windows. Given that these data sets had been created with $N = 4{,}000$, one might have expected that using just 50 windows, with $n = 678$, would be insufficient to identify categorical structure with as few as $n = 4{,}000 \times .05 = 200$ taxon members in the sample. Nonetheless, CCFI results unambiguously identified categorical structure in each instance without recourse to the inchworm consistency test. Whereas the subjective, visual inspection of curves may be facilitated by attempts to determine whether or not a cusp becomes a better defined peak as the number of windows increases, the CCFI does not require that a peak emerge to identify categorical structure. In fact, a cusp may not be ambiguous when considered in light of results for comparison data. Thus, the CCFI can be effective even with a few windows and a small taxon. Future research should examine the incremental validity of performing a series of analyses with an increasing number of windows relative to a single analysis with comparatively few windows.

In extending the CCFI to low base rate constructs, it should be noted that in the current Monte Carlo investigation (Study 1) P never fell below .10. One limitation of this study, in fact, is that challenging or unfavorable parameters (e.g., $N < 300$, $d < 1.25$, within-group $r > .30$, $P < .10$) were not systematically evaluated. When Ruscio et al. (in press) examined challenging and unfavorable parameters in a large-scale Monte Carlo analysis, the CCFI displayed a sharp decline in accuracy for taxonic samples with below-threshold indicator validity ($d < 1.25$) or above-threshold nuisance covariance ($r > .30$), whereas the accuracy of the CCFI in taxonic samples with below-threshold base rates ($.05 < P < .10$) declined modestly and taxonic and dimensional samples with below-threshold sample sizes ($100 \leq N \leq 300$) displayed slight and sharp decrements in accuracy, respectively. Accuracy tends to decline even more when several parameters fall below threshold, although the CCFI is reasonably robust in the face of multiple parameters slightly above threshold (Ruscio & Walters, 2009). The goal of the present research was to examine the effect of certain implementation decisions on the accuracy level of the CCFI rather than evaluate the absolute accuracy of taxometric analysis under various data conditions. Additional research will be required to determine the extent to which CCFI accuracy is affected by a very low base rate ($P < .05$), in which case it might prove helpful to increase the number of cuts in MAMBAC beyond 50 and the number of windows in MAXCOV/MAXEIG beyond 25, or an increase in the number of outliers in excess of what occurs by normal sampling error (and which was therefore incorporated in our Monte Carlo study).

Within the range of data conditions for which taxometric procedures are expected to provide informative results (Meehl, 1995), the present findings suggest the most effective default values for implementing the MAMBAC, MAXCOV, and MAXEIG taxometric procedures. It is possible that alternative implementations might be more effective for other data conditions, but we chose not to investigate performance beyond the limits for which taxometric analysis is recommended (e.g., very small group separation)—or for which it may not be necessary (e.g., very large group separation). There is at least one additional data parameter that merits further study, however: the analysis of ordered categorical data. Our simulation study (Study 1) included truly continuous variables, but researchers often have available only ordered-categorical variables (e.g., binary items, Likert-type scales, or composite scores formed from these ordered-categorical item types). We examined indicators composed of a modest number of ordered categories (5 to 7) in Study 2 and obtained results that paralleled findings from our Monte Carlo study. Of course, these results are merely suggestive given that they were based on a single sample. Walters and Ruscio (2009) found that with ordered categories, the performance of taxometric analyses declined sharply with fewer than four categories per variable and fewer than five variables. Their MAMBAC analyses were performed with $n = 25$ cases reserved beyond the first and last of 50 cuts, and their MAXCOV and MAXEIG analyses were performed with 50 windows. It would be interesting to extend those results across other implementations of each procedure.[3] In the meantime, reserving $n = 10$ cases beyond the first and last cuts in MAMBAC analyses and performing MAXCOV or MAXEIG analyses with 25 windows

appears to provide the most accurate results. This constitutes the first set of empirically derived guidelines for implementing these procedures, and the scope of the evidence on which they are based suggests that they can serve as a preliminary framework for making certain taxometric implementation decisions until new data refine or refute them.

Authors' Note

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.

Acknowledgments

The authors would like to thank Robert Day for running half of the samples included in Study 1.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Notes

1. In the literature on taxometrics, the higher scoring of two groups is usually referred to as the "taxon," and the other group the "complement."
2. The error (miss) rate for the three most accurate procedures (MAMBAC-10, MAXCOV-25W, MAXEIG-25W) was 16 times higher for the comparison curve fit index (CCFI) values falling within the narrow dual threshold range (36.3% to 40.2%, $M = 38.1\%$) than for CCFI values falling outside the narrow dual threshold range (1.5% to 2.9%, $M = 2.4\%$). The error rate for these same three procedures was 22 times higher for CCFI values falling within the broad dual threshold range (23.2% to 27.9%, $M = 26.0\%$) than for CCFI values falling outside the broad dual threshold range (0.5% to 1.8%, $M = 1.2\%$).
3. We were unable to do so in the present study because of computational demands. For each of our 10,000 target data sets, a total of 29 taxometric analyses was performed, each consisting of a panel of results to calculate and each with its own parallel analyses of categorical and dimensional comparison data. This took a considerable amount of time to run. Had we studied ordered categorical data instead, this would have necessitated the use of internal replications (Ruscio et al., 2006) to reduce the influence of tied scores. Previous research has used 10 internal replications, and this would have increased our run time by approximately a factor of 10, rendering the study infeasible given our present computational resources.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Dana, D. A. (1990). Alcoholism and antisocial personality disorder: Related discrete personality variables? *Dissertation Abstracts International, 51*(04), 2056B. (University Microfilms No. AAT90-24616)

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

Green, B. A., Ahmed, A. O., Marcus, D. K., & Walters, G. D. (2009). *The latent structure of alcohol use pathology in an epidemiological sample.* Manuscript submitted for publication.

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.

Lenzenweger, M. F. (2004). Consideration of the challenges, complications, and pitfalls of taxometric analysis. *Journal of Abnormal Psychology, 113*, 10-23.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50*, 266-275.

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

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

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

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

Ruscio, J. (2009). *Taxometric programs for the R computing environment: User's manual* [computer software and manual]. Retrieved November 11, 2009, from http://www.taxometricmethod.com

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method: A practical guide.* Mahwah, NJ: Lawrence 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 analyses. *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., & Keane, T. M. (2004). Using taxometric analysis to distinguish a small taxon from a latent dimension with positively skewed indicators: The case of involuntary defeat syndrome. *Journal of Abnormal Psychology, 113*, 145-154.

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.

Ruscio, J., Walters, G. D., Marcus, D. K., & Kaczetow, W. (in press). Comparing the relative fit of categorical and dimensional latent variable models using consistency tests. *Psychological Assessment*.

Slade, T., Grove, R., & Teeson, M. (2009). A taxometric analysis of alcohol abuse and dependence in a general population sample: Evidence of dimensional latent structure and implications for *DSM-V*. *Addiction, 104*, 742-751.

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

Walters, G. D. (2008). The latent structure of alcohol use disorders: A taxometric analysis of structured interview data obtained from male federal prison inmates. *Alcohol and Alcoholism, 43*, 326-333.

Walters, G. D. (2009). Taxometric analysis of alcohol dependence in male prisoners: Measuring latent structure with indicators from *DSM-IV*. *Addiction Research & Theory, 17*, 372-380.

Walters, G. D., Diamond, P. M., & Magaletta, P. R. (in press). What is the latent structure of alcohol use disorders? A taxometric analysis of the Personality Assessment Inventory (PAI) alcohol problems scale in male and female inmates. *Psychology of Addictive Behaviors*.

Walters, G. D., Henning, C. L., Negola, T. D., & Fricke, L. A. (2009). The latent structure of alcohol dependence in female federal prisoners. *Addiction Research & Theory, 17*, 525-537.

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