Assessment
Volume 16 Number 1
March 2009 55-70

10.1177/1073191108320193
http://asmnt.sagepub.com
hosted at
http://online.sagepub.com

Assigning Cases to Groups Using Taxometric Results

An Empirical Comparison of Classification Techniques

John Ruscio
The College of New Jersey

Determining whether individuals belong to different latent classes (taxa) or vary along one or more latent factors (dimensions) has implications for assessment. For example, no instrument can simultaneously maximize the efficiency of categorical and continuous measurement. Methods such as taxometric analysis can test the relative fit of taxonic and dimensional models, but it is not clear how best to assign individuals to groups using taxometric results. The present study compares the performance of two classification techniques—Bayes' theorem and a base-rate technique—across a wide range of data conditions. The base-rate technique achieves greater classification accuracy and a more even balance between sensitivity and specificity. In addition, the base-rate classification technique is easier to implement than Bayes' theorem and is more versatile in that it can be used when the context of assessment requires that cases be classified despite the absence of latent classes.

Keywords: *classification; taxometrics; Bayes' theorem; base rate; categories; latent structure*

Whether a construct is categorical or continuous has important implications for its assessment (Meehl, 1992; Ruscio & Ruscio, 2002). For example, an instrument cannot simultaneously maximize the efficiency of continuous and categorical measurement. When individuals vary along one or more latent continua, researchers or practitioners ordinarily attempt to locate examinees' positions using measurement instruments that provide continuous or quasicontinuous scores. Psychometric techniques for working with continuous scores as estimates of scores on latent factors are familiar to psychologists. When individuals vary across discrete groups, a more appropriate goal for psychological assessment often is to assign examinees to groups. Methods for determining when the best-fitting structure includes latent classes and how to use available data to classify cases are less familiar to many psychologists. Ruscio and Ruscio discussed some of the assessment-related issues raised by the distinction between latent categories and continua and provided an overview of Meehl's (1995) taxometric procedures for making this distinction empirically. The present study extends this earlier work by focusing on the question of how to assign cases to groups using the results of a taxometric analysis.

Meehl's (1995) taxometric method can be used to distinguish taxonic (categorical) and dimensional (continuous) latent variables. The method contains a number of data-analytic procedures that can be used to evaluate the consistency with which results support an inference of taxonic or dimensional latent structure (e.g., Grove, 2004; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998). Evidence suggests that taxometric analyses validly differentiate taxonic and dimensional latent variables under a wide range of data conditions, at least when comparison data are used as an interpretive aid (Ruscio, 2007; Ruscio & Marcus, 2007; Ruscio, Ruscio, & Meron, 2007). Ruscio, Haslam, and Ruscio (2006) provided an introduction to the method that describes the difference between taxonic and dimensional latent structures, the taxometric procedures and consistency tests most frequently used in research, and what is known about their performance; briefer taxometric tutorials are also available (e.g., Ruscio, 2007; Ruscio & Ruscio, 2004a, 2004b).

After performing a taxometric analysis, often it is useful to assign cases to groups for subsequent analysis or for practical purposes. For example, one might

Author's Note: The author is grateful to David Marcus and Walter Kaczetow for providing helpful comments on earlier drafts of this work. Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628; e-mail: ruscio@tcnj.edu.

wish to compare groups assigned to different diagnostic categories. The group whose members score higher on the available measures is referred to as the *taxon* (e.g., individuals diagnosable with Major Depressive Disorder [MDD]; MDD+) and the lower-scoring group the *complement* (e.g., individuals not diagnosable with MDD; MDD–). If the group of primary interest scores lower on the available measures (e.g., MDD scales for which higher scores represent more adaptive cognitive, affective, or behavioral responses), typically each variable is reverse-scored prior to analysis so that the taxon becomes the higher-scoring group. Naturally, one would like to use the observed measures to assign cases to groups in a way that corresponds most closely to their membership in the latent taxon and complement classes. Any technique will be fallible, as some MDD+ individuals (taxon members) will score lower than some MDD– individuals (complement members) on the measures in a given study due to imperfect validity and measurement error. There are at least two ways to use taxometric results to classify cases into groups.

First, one can use estimates of latent parameters of the taxonic structural model to calculate each individual's probability of taxon membership with Bayes' theorem. This requires estimates of the taxon base rate (P, the proportion of MDD+ individuals in the sample), the complement base rate ($Q = 1–P$, the proportion of MDD– individuals), and the true and false positive rates achieved by the optimal cutting score on each of k indicator variables ($TP_1 \ldots TP_k$ and $FP_1 \ldots FP_k$; a variable's TP rate is the estimated proportion of above-threshold scores that belong to MDD+ individuals, and the FP rate is the estimated proportion of above-threshold scores that belong to MDD– individuals). These estimates can be obtained from the results of a taxometric analysis (Ruscio et al., 2006). The following formula yields the probability of taxon membership for each case (Meehl & Yonce, 1996):

$$\Pr(taxon) = \frac{P\prod_{j=1}^{k} TP_j^{\theta}(1-TP_j)^{1-\theta}}{P\prod_{j=1}^{k} TP_j^{\theta}(1-TP_j)^{1-\theta} + Q\prod_{j=1}^{k} FP_j^{\theta}(1-FP_j)^{1-\theta}},$$

where Π is the cumulative product operator and $\theta = 1$ for an above-threshold response, 0 for a below-threshold response. Using this formula, one can estimate Pr(*taxon*) for each case (e.g., the probability that this individual is MDD+). Because Pr(*complement*) = 1 – Pr(*taxon*) for each individual, it is a simple matter to assign each individual to the more probable group (i.e., assign to the MDD+ group if Pr(*taxon*) > .5, otherwise assign to the MDD– group). In this article, the notation C_{Bayes} represents the use of Bayes' theorem to classify cases.

A second method is to sort cases based on their total scores on the available indicator variables (e.g., MDD symptom ratings) and assign individuals scoring highest to the taxon. A threshold is used to assign to the taxon a proportion of cases equal to the estimated taxon base rate (Ruscio et al., 2006). For example, if the estimated MDD+ base rate is .30 in the population from which the sample was drawn, then the highest-scoring 30% of cases are assigned to the MDD+ group and the remaining 70% of cases are assigned to the MDD– group. The notation C_{P} represents this base-rate method for classifying cases. A practical advantage of C_{P} is that it requires only a base-rate estimate, which might be obtained in any number of ways (e.g., taxometric analysis, well-supported theory, prior research findings). In contrast, C_{Bayes} requires not only a base-rate estimate, but also estimates of the TP and FP rates in the appropriate population for each variable at hand. These estimates are more challenging to obtain, and doing so in a taxometric study constrains the options for implementing taxometric procedures (Ruscio et al., 2006). Because C_{P} is simpler and requires the estimation of fewer parameters than C_{Bayes}, it may be appealing in a broader range of circumstances. A more important consideration, however, is classification accuracy.

Meehl (1973), Meehl and Golden (1982), and Meehl and Yonce (1996) demonstrated that C_{Bayes} can achieve impressive accuracy. The generalizability of these findings is uncertain, however, because the idealized data conditions in these studies may not represent adequately the complexities inherent in actual data sets. For example, indicator variables in these studies were normally distributed along continuous scales, whereas most research data are not (Micceri, 1989): Skewed distributions along ordered categorical (e.g., Likert-type) scales are common. Beauchaine and Beauchaine (2002) varied data conditions along a greater number of factors and a broader range of levels, and they compared the accuracy achieved by C_{Bayes} to that of k-means cluster analysis. They found that C_{Bayes} surpassed the accuracy of the k-means technique only under certain conditions in which a number of data parameters simultaneously departed from idealized values. This suggests that C_{Bayes} may be useful with actual data, but it does not afford a comparison with other techniques that might be applied using the same results of a taxometric analysis.

Ruscio et al. (2006, chap. 6) performed a small study that compared C_{Bayes} and C_P. Each taxonic data set included $k = 4$ indicators in a sample of $N = 1{,}000$ cases. Within each group, indicators were normally distributed along continuous scales. Three data parameters were crossed in this simulation study: taxon base rate (P) = .50, .25, or .10; indicator validity (between-group separation, indexed using Cohen's d) = 2.00, 1.50, or 1.00; and within-group correlation (r) = .00, .25, or .50. For each of the 27 cells in this design, 100 samples were generated using a data simulation technique described by Meehl and Yonce (1994, 1996). Cases were classified using C_P with either the true taxon base rate or a value that erred by ±10%, ±25%, or ±50%. Increasingly poor base-rate estimates were used to examine the robustness of C_P to such misspecifications. The benchmark for comparison was quite stringent: Cases were classified using C_{Bayes} with the actual (rather than estimated) parameters of the taxonic structural model. For both methods, classification accuracy was measured as the proportion of all cases assigned to the correct group, or the hit rate (*HR*). C_P outperformed C_{Bayes} in 26 out of 27 data conditions (96%) when provided with the correct base rate and 91% of conditions when provided with base rates that erred by ±10%. When provided with base rates that erred by ±25% and ±50%, C_{Bayes} was more accurate in 51% and 89% of data conditions, respectively. This suggests that when one has a reasonably accurate estimate of the taxon base rate, C_P can classify cases with comparable or superior accuracy to C_{Bayes}.

Although these results support the potential utility of C_P, the design of this small-scale study can be improved in a number of ways. First, it was a very conservative test of C_P because its competitor (C_{Bayes}) was provided with infallible parameter values. In actual research, only data-based parameter estimates will be available for any classification method. Second, neither method had been evaluated using the results of taxometric analyses. Researchers often perform taxometric analyses to estimate parameters of the taxonic structural model, and it would be informative to know how accurately cases can be classified using these estimates. The present study addressed these first two limitations by comparing the C_{Bayes} and C_P methods when each was performed using the results of the same taxometric analyses.

Third, the range of data conditions in the small-scale study was rather limited. For example, in actual data sets, indicators seldom will be normally distributed, and data often vary across ordered categorical response scales rather than truly continuous scales. Whereas the previous investigation included data conditions that varied along three factors (taxon base rate, indicator validity, and within-group correlations), the present study included data conditions that varied along these factors plus four others (sample size, number of indicators, indicator skew, and continuous vs. ordered categorical response scales of varying sizes). This provides much broader coverage of the parameter space likely to be encountered by researchers.

Fourth, there is more than one way to calculate a taxon base-rate estimate for C_P. One can use the mean base-rate estimate from a series of taxometric analyses. For example, analyses may yield a mean estimate .25, in which case C_P would assign the highest-scoring 25% of the sample to the MDD+ group. When C_P uses a taxometric base-rate estimate, this will be referred to as $C_{P\text{-}T}$. Alternatively, one can classify cases using C_{Bayes} and then compute the proportion of cases assigned to the taxon as a new estimate of the taxon base rate. For example, using the parameter estimates calculated from taxometric results, Bayes' theorem may assign 35% of cases to the MDD+ group, in which case C_P would assign the highest-scoring 35% of the sample to the MDD+ group (rather than the 25% suggested by the taxometric base-rate estimate). It may be the case that the proportion of cases assigned to the taxon using Bayes' theorem equals the original taxometric base-rate estimate, in which case the use of C_P may appear superfluous or redundant. However, even when C_{Bayes} and C_P assign the same proportion of cases to a taxon, the specific cases in the taxa can—and usually will—differ.[1] When C_P uses the reestimated base rate calculated after a preliminary classification using Bayes' theorem, this will be referred to as $C_{P\text{-}B}$. Finally, the base-rate estimates used by $C_{P\text{-}T}$ and $C_{P\text{-}B}$ can be averaged. For example, if $C_{P\text{-}T}$ and $C_{P\text{-}B}$ assign 25% and 35% of cases to the MDD+ group, respectively, using the average of these two base-rate estimates would assign the highest-scoring 30% of cases to the MDD+ group. When C_P uses this mean base-rate estimate, this will be referred to as $C_{P\text{-}M}$. To recap, each of these three versions of C_P begins by sorting cases according to their total scores on the available indicators, with the difference then being what proportion of the highest-scoring cases are assigned to the taxon. To the extent that the base rates estimated by the original taxometric analysis and the Bayesian classification converge, so too will the thresholds used by these three versions of C_P.

Fifth, using *HR* as the measure of classification accuracy does not control for base rates or chance-level agreement. When the taxon base rate is very

small (or large), any classification technique can achieve a high *HR* by assigning most or all cases to the larger group. For example, if 70% of cases belong to an MDD+ group, blindly assigning everyone to this group would achieve *HR* = .70. Likewise, even randomly assigning 70% of the cases to the MDD+ group would be expected to yield a substantial *HR* of .70 × .70 + .30 × .30 = .58 by chance alone. The present study included an adjusted *HR* that corrects for chance as well as another measure that is independent of base rates and corrects for chance.

With each of these five concerns in mind, the present study was designed to afford a more rigorous test of the relative classification accuracy of C_{Bayes} and C_P, including several ways to implement the latter. The goal was to provide researchers with guidance on how to assign cases to groups for subsequent analysis or for practical purposes.

Method

Design and Data Generation

Rather than basing data conditions on those in specific empirical data sets, the present study was designed to cover a sufficiently broad range that it should span the realistic research conditions one might encounter, and then some. A total of 100,000 taxonic data sets were generated using the same Monte Carlo design employed by Ruscio et al. (2007, Study 3) and Ruscio (2007, Study 2). Data parameters were independently randomly sampled from specified ranges that broadly spanned values considered to be adequate for informative taxometric analyses according to rules of thumb provided by Meehl (1995). Although in many ways similar to the study design used by Beauchaine and Beauchaine (2002), there are three noteworthy differences. First, the present study only included values along each factor within the ranges recommended for taxometric analysis by Meehl. Beauchaine and Beauchaine included values that extended beyond these ranges to test Meehl's rules of thumb.

Second, whereas Beauchaine and Beauchaine's (2002) indicator distributions were normal in shape and continuous, the present study also included skewed and ordered categorical distributions. Micceri (1989) demonstrated that empirical data sets seldom exhibit normal distributions, and ordered categorical response scales are common as well.

Third, the present study used a crossed factorial design. Beauchaine and Beauchaine drew samples at fixed intervals along one factor in their design (e.g., N = 1,000, 900, 800, . . .), with values on all other factors held constant at the most favorable levels. Their final series of analyses incrementally incorporated more challenging values for multiple factors, but this represented one unique path through the possible parameter space. In the present study, a crossed factorial design was used in which each target data set consisted of a set of randomly chosen values for each parameter in the design. This yields a representative sample of all possible paths through the parameter space.

For each target data set in the present study, values were drawn for the following parameters: sample size (N = 300 to 1,000), taxon base rate (P = .05 to .50), indicator validity (d = 1.25 to 2.00), within-group correlation (r = .00 to .30), indicator skew (S = 0, 1, 2, . . . , 6), number of indicators (k = 3, 4, 5, . . . , 8), and number of ordered categories (C = 6, 9, 12, 15, 0, with 0 representing continuous distributions and all other values representing the number of response options—i.e., 6-point Likert-type scales, 9-point Likert-type scales, etc.). Values of N, P, d, and r were drawn from uniform, continuous distributions spanning the ranges listed above. With one exception ($P \geq .05$ rather than $P \geq .10$), these represent values acceptable for taxometric analysis according to Meehl's (1995) rules of thumb. Slightly lower base rates were included because researchers often investigate rare phenomena and it seemed important to evaluate classification accuracy under these conditions. For the remaining factors, no rules of thumb were available to guide the selection of parameter values, and so ranges were designed to span values that researchers might encounter, including ideal and suboptimal levels. Values of k and C were drawn from uniform distributions spanning the categories listed above. Values of S were drawn from the categories listed above, but S was determined at random with the following probabilities: 0 (.25), 1 (.20), 2 (.20), 3 (.15), 4 (.10), 5 (.05), and 6 (.05). Because Micceri (1989) classified 28.4% of the 440 empirical data sets that he reviewed as "relatively symmetric," 40.7% as "moderate asymmetry," 19.5% as "extreme asymmetry," and 11.4% as "exponential asymmetry," the present study was designed such that more extreme skew levels occurred with lower frequencies.

Data were generated within each group (taxon and complement) according to the group size ($N \times P$ for the taxon, $N \times [1 - P]$ for the complement) and values of r, S, k, and C randomly chosen for that sample using the GenData program shown in the appendix. This program uses an iterative approach to generating

nonnormal multivariate data with specified univariate distributions and indicator correlations based on Ruscio and Kaczetow's (in press) generalization of the DimSample program that was originally developed by Ruscio et al. (2007) and is used often in taxometric investigations. Ruscio and Kaczetow demonstrated the broad utility of this iterative approach to generating multivariate nonnormal data, and it was implemented in the present study using normal distributions (when $S = 0$) or lognormal distributions (when $S > 0$). By using the same values of r, S, k, and C for each group, indicator distributions and correlation matrices were held constant across groups; this was done for simplicity of design, with no expectation that it should give an edge to C_{Bayes} or C_{P}. As shown in the program code in the appendix, the data generation model presumes local independence, that the model holds within both groups, that within-group correlations are due to a single latent factor, and that the error of measurement is constant. Data were standardized on each indicator within each group, a constant equal to d was added to all scores in the taxon, and the taxon and complement subsamples were merged to yield a taxonic data set.

Data Analysis and Classification

Each sample was analyzed using the MAXimum EIGenvalue (MAXEIG; Waller & Meehl, 1998) taxometric procedure, a multivariate extension of the MAXimum COVariance (MAXCOV; Meehl & Yonce, 1996) procedure. Whereas all taxometric procedures provide one or more estimates of the taxon base rate, MAXEIG and MAXCOV are the only procedures in common usage that provide estimates of the true and false positive rates for each indicator, which are required to implement C_{Bayes}. There are no analytical or empirical comparisons of the accuracy with which these procedures estimate all of these parameters, and therefore no evidence-based reason to prefer either procedure. MAXEIG was used for analytic convenience, as it required considerably fewer analyses than MAXCOV and therefore facilitated the inclusion of a larger number of target data sets.[2] MAXEIG was performed such that adjacent subsamples (windows) overlapped 90% with one another ($O = .90$). For the analysis of each target data set, the number of windows (W) was determined such that there would be at least $n_w = 50$ cases within each window. This was done by solving a formula provided by Waller and Meehl (1998, p. 42) using each sample's N, holding constant $n_w = 50$ and $O = .90$, and dropping any decimal:

$$W = \frac{\frac{N}{n_w} - O}{1 - O}.$$

Thus, W ranged from 51 (when $N = 300$) to 191 (when $N = 1{,}000$). For samples with ordered categorical data (i.e., when $C > 0$), results were averaged across 10 internal replications to reduce the obfuscating influence of assigning cases with tied scores to different subsamples (Ruscio et al., 2006).

From these MAXEIG results, parameters of the taxonic structural model were estimated using the technique described in Ruscio et al. (2006, Appendix C). Using these estimates, cases were assigned to groups with each method described earlier: C_{Bayes}, $C_{\text{P–T}}$, $C_{\text{P–B}}$, and $C_{\text{P–M}}$.

Evaluating the Classifications

The first step was to calculate sensitivity and specificity (Bossuyt et al., 2003; Streiner, 2003). Sensitivity was calculated as the proportion of actual taxon members correctly classified, and specificity was calculated as the proportion of actual complement members correctly classified. For each classification, the resulting sensitivity and specificity values can be used to generate a binormal receiver operating characteristic (ROC) curve, and the area beneath this curve (A) provides an excellent measure of overall classification accuracy (Swets, 1988). A is independent of base rates and ranges from a maximum of 1.00 (perfect accuracy) through 0.50 (chance-level accuracy) to a minimum of .00, with values < .50 representing accuracy worse than chance (i.e., A would be higher if each case's classification was reversed).

To demonstrate that findings were not dependent on the choice of A as a measure of classification accuracy, HR was calculated as the proportion of all cases correctly classified. Because, as noted earlier, HR does not take into account the hits expected due to chance (e.g., it is easier to classify a large proportion of cases correctly when one group is very large and the other very small than when the groups are of more equal size), it was corrected for chance using the formula for Cohen's κ (Cohen, 1960):

$$HR_{\text{adj}} = \frac{HR - CA}{1 - CA},$$

Table 1
Summary of Results for 100,000 Samples

Classification Technique	Sensitivity	Specificity	*A*	*R*	*HR*	HR_{adj}
C_{Bayes}	.796	.913	.939	.466	.869	.599
C_{P-T}	.853	.994	.994	.462	.929	.787
C_{P-B}	.976	.966	.996	.502	.940	.820
C_{P-M}	.926	.986	.995	.484	.940	.819

Note: Sensitivity, specificity, *HR* (hit rate), and HR_{adj} (adjusted hit rate) values were averaged across samples using 20% trimmed means to increase robustness to the influence of outliers in each distribution (Wilcox, 2003). *A* = area under receiver operating characteristic curve; *R* = ratio of sensitivity to sum of sensitivity and specificity; C_{Bayes} = classification using Bayes' theorem; C_{P-T} = base-rate classification using the mean base-rate estimate from MAXEIG (MAXimum EIGenvalue) analyses; C_{P-B} = base-rate classification using the proportion of cases assigned to the taxon using Bayes' theorem; C_{P-M} = base-rate classification using the mean of the base-rate estimates of the previous two techniques.

where HR_{adj} is the adjusted hit rate, *HR* the observed hit rate, and *CA* the expected accuracy due to chance for an actual base-rate π and base-rate estimate *P*, calculated as $CA = (P)(\pi) + (1 - P)(1 - \pi)$. For example, suppose that the actual base rate of MDD+ is $\pi = .20$, the estimate is $P = .25$, and the observed *HR* is .80. Correcting for $CA = (.25)(.20) + (.75)(.80) = .65$ yields $HR_{adj} = (.80 - .65)/(1 - .65) = .43$. Like Cohen's κ, this ranges from a maximum of 1.00 (perfect accuracy) through 0.00 (chance-level accuracy) to negative values that represent accuracy worse than chance; mathematically, there is no lower limit to this scale.

In addition to overall classification accuracy, the balance between sensitivity and specificity was examined. All else being equal, a method that misclassifies members of one group at a higher rate than members of another group is less desirable than a method that classifies members of both groups with comparable accuracy levels. The ratio *R*, defined as follows, was calculated to index this balance:

$$R = \text{sensitivity}/(\text{sensitivity} + \text{specificity}).$$

Values of *R* could range from a minimum of 0.00 (when sensitivity = 0) to a maximum of 1.00 (when specificity = 0), with $R = .50$ obtained when sensitivity equals specificity.

Results

Overall Summary

Table 1 summarizes the results across all 100,000 data sets. Each method achieved an accuracy far in excess of chance: C_{Bayes} yielded $A = .939$ and C_P yielded *A*s from .994 to .996 depending on which base rate was provided. C_{Bayes} also achieved a lower *HR* (.869) and HR_{adj} (.599) than C_P (*HR* from .929 to .940, HR_{adj} from .787 to .820). Although both techniques achieved high levels of accuracy, C_P yielded larger values of both sensitivity (.853 to .976) and specificity (.966 to .994) than C_{Bayes} (sensitivity = .796, specificity = .913).

C_{Bayes} achieved greater specificity than sensitivity ($R = .466$). Among the three types of C_P, the use of C_{P-B} achieved the best balance between sensitivity and specificity ($R = .502$); using either C_{P-T} or C_{P-M} yielded greater specificity than sensitivity ($R = .462$ and .484, respectively). Given that the taxon base rate was $\leq .50$ for all samples in this study, it should not be surprising that members of the smaller taxon were usually more difficult to identify than members of the larger complement.

The variation in *R* values for the three types of C_P is attributable to the differential accuracy of the taxon base-rate estimates that were used to classify cases. Across all samples, the base-rate estimate for C_{P-B} was most accurate, with a mean discrepancy (actual – estimated *P*) of .013. The base-rate estimate for C_{P-T} tended to be too low (mean discrepancy = –.049), and even though that for C_{P-M} underestimated by a lower amount (mean discrepancy = –.018), it was still less accurate than that for C_{P-B}.

With relatively few exceptions, the rank-ordering of each technique's accuracy, balance of sensitivity and specificity, and base-rate discrepancies for the full set of 100,000 samples was observed consistently across data conditions. Specific results for each of these outcome measures are discussed in the next three subsections. Within data conditions, only the average results are presented because the large number of target data sets yielded very small standard errors. The smoothness of the curves in the figures cited below attests to the negligible magnitude of sampling error. It should be noted,

however, that occasional outliers on the outcome measures yielded negatively skewed distributions. When sensitivity, specificity, *HR*, and HR_{adj} values were averaged within data conditions, 20% trimmed means were used to increase robustness to the influence of outliers (Wilcox, 2003). *A* and *R* values were calculated using the trimmed-mean sensitivity and specificity values.

Classification Accuracy Across Data Conditions

To simplify presentation, only results for *A* are reported; similar patterns of findings emerged for *HR* and HR_{adj}. Although classification accuracy varied across data conditions (see Figure 1)—especially for C_{Bayes}—all three types of C_P attained greater accuracy than C_{Bayes} under all conditions. Each technique's accuracy was at least as good with smaller as with larger samples and improved with larger numbers of indicators, higher indicator validity, lower within-group correlations, and data that better approximated continuous score distributions (i.e., with larger numbers of ordered categories if not truly continuous). The results for taxon base rate and indicator skew were a bit more complex. C_{Bayes} was less accurate with lower base rates, whereas C_P was most accurate with low base rates, a little less accurate with base rates in the middle of the range studied here, and increasingly accurate with base rates at the upper end of the range. This latter pattern was also observed for the accuracy of C_P across levels of indicator skew, whereas the accuracy of C_{Bayes} declined with increasing indicator skew.

The three variants of C_P usually achieved accuracy levels comparable to one another, with nontrivial exceptions in only two conditions. First, at very high levels of indicator skew ($S \geq 3$), the use of $C_{P\text{-}B}$ achieved the poorest accuracy. Second, when indicators varied across smaller numbers of ordered categories ($C = 6$ or 9), the use of $C_{P\text{-}B}$ achieved the greatest accuracy.

Taken as a whole, the results for classification accuracy suggest that C_P should be preferred over C_{Bayes}, and that $C_{P\text{-}B}$ nearly always achieved accuracy levels comparable or superior to $C_{P\text{-}T}$ or $C_{P\text{-}M}$. Not only was accuracy greater for C_P than C_{Bayes} across the range of values included for each factor in the design, but the gap widened as data conditions became less favorable for taxometric analysis (e.g., smaller indicator validity, larger within-group correlations). This suggests that extending the ranges to even less favorable conditions (e.g., $d < 1.25$, $r > .30$) would provide even stronger support for C_P.

Sensitivity and Specificity Across Data Conditions

Figure 2 shows the balance between sensitivity and specificity across data conditions. Consistent with the global results presented earlier, sensitivity was usually lower than specificity ($R < .50$). This was fairly constant across sample sizes and indicator validities, but *R* did increase with larger numbers of indicators and larger within-group correlations. C_{Bayes} yielded larger *R* values with larger taxon base rates, whereas the opposite was true for C_P. Increasing indicator skew substantially decreased *R*. C_{Bayes} yielded lower *R* values as indicator score distributions better approximated continuity, whereas C_P yielded *R* values that approached .50 as distributions better approximated continuity.

Perhaps most important is that among the four classifications studied, $C_{P\text{-}B}$ yielded a value of *R* closest to .50 for most conditions. There were two notable exceptions. First, when the taxon base rate was low ($P < .15$), the other two types of C_P yielded values of *R* closer to .50; $C_{P\text{-}B}$ yielded *R* values well below .50. Second, when indicators were not skewed ($S = 0$), all other classifications yielded *R* values closer to .50 than $C_{P\text{-}B}$, and when indicators were mildly skewed ($S = 1$), $C_{P\text{-}M}$ still achieved an *R* closer to .50 than $C_{P\text{-}B}$. These results suggest that $C_{P\text{-}B}$ most often provides a balance between sensitivity and specificity comparable or superior to the other techniques studied.

Base-Rate Estimation Across Data Conditions

Figure 3 shows discrepancies in base-rate estimation across data conditions. The mean value of estimates for $C_{P\text{-}B}$ was the largest, and the mean value of estimates for $C_{P\text{-}T}$ was the smallest (estimates for $C_{P\text{-}M}$, of course, fell at an intermediate level because they were calculated as the mean of the other two). As was observed across all samples, estimates for $C_{P\text{-}T}$ (and, to a lesser extent, $C_{P\text{-}M}$) tended to underestimate base rates more than those for $C_{P\text{-}B}$ tended to overestimate them. This suggests that in many circumstances, researchers may be well-advised to report and interpret the estimate for $C_{P\text{-}B}$ as the best estimate of the taxon base rate, rather than the conventional estimate for $C_{P\text{-}T}$. However, discrepancies in base-rate estimation varied to some extent with variation along each data parameter. By far, the most substantial effects were observed across the actual taxon base rate and indicator skew. Base rates were overestimated at small values of *P* or *S* and underestimated at

Figure 1
Accuracy of Classification Across Data Conditions, Quantified as the Area Under a Receiver Operating Characteristic (ROC) Curve

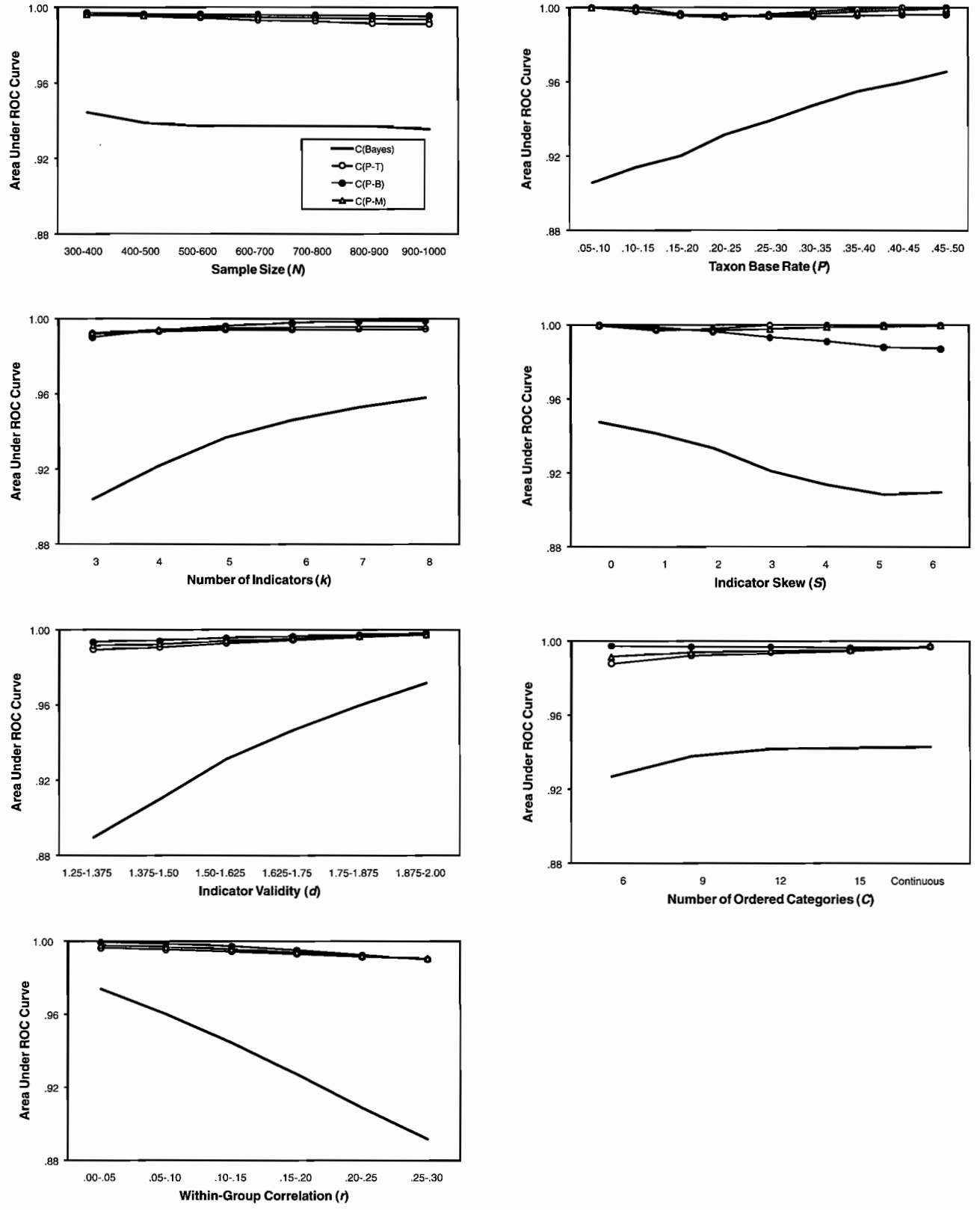
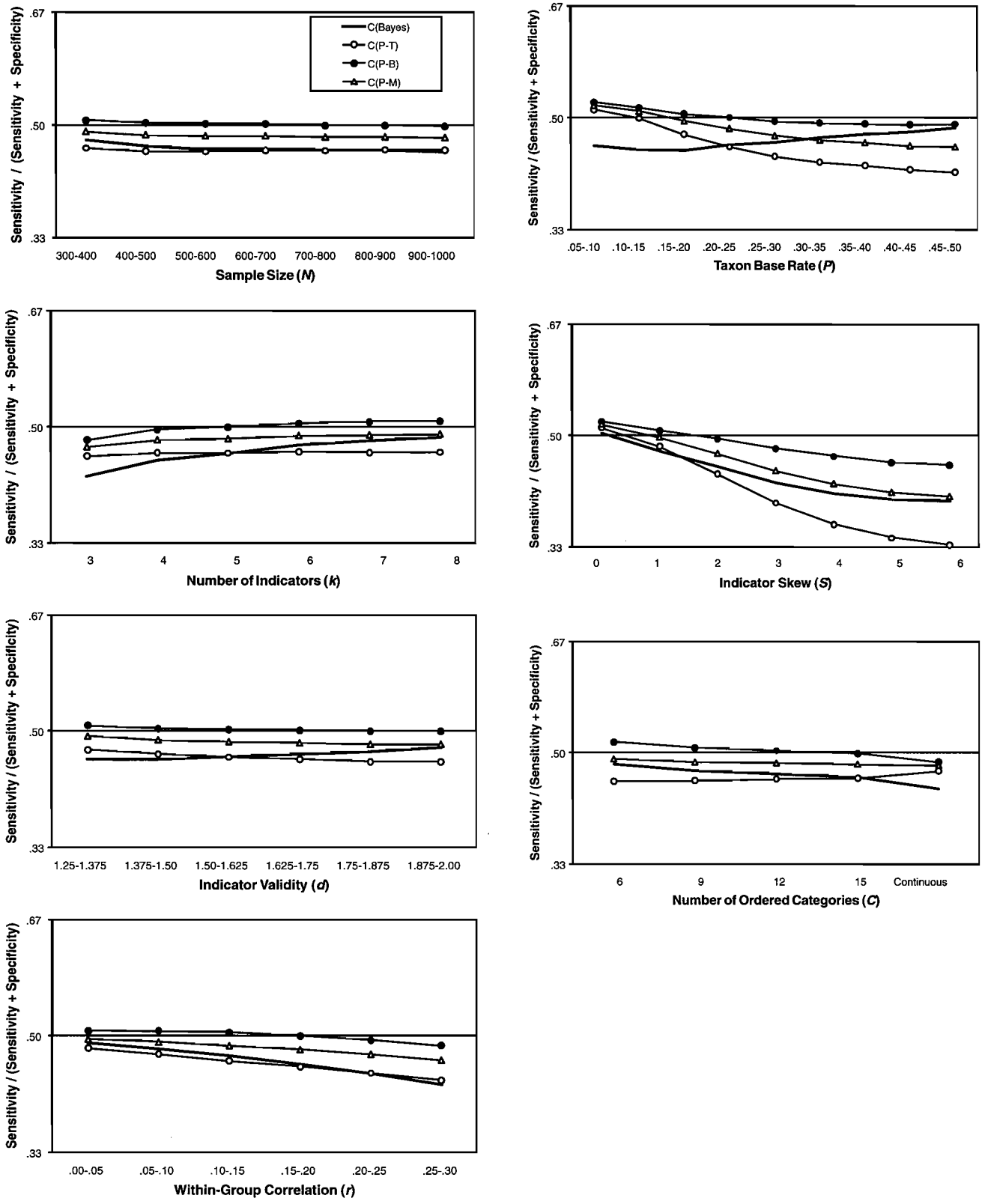


Figure 2
Balance of Sensitivity and Specificity Across Data Conditions



Note: Dotted lines at y = .50 represent equal values for sensitivity and specificity.

high values. The magnitude of these discrepancies is striking, and even more so when the joint effects of *P* and *S* are examined. For example, out of the 100,000 samples generated in this study, 2,767 satisfied the conditions that $P < .10$ and $S = 0$. Their actual base rates averaged a value of $M = .075$, and this was overestimated by an average of .167 to .215, depending on which C_P method was used. At the other end of this spectrum, 541 samples satisfied the conditions that $P > .45$ and $S = 6$. Their actual base rates averaged $M = .474$, and this was underestimated by an average of .172 to .333. The graph in the lower right corner of Figure 3 elaborates by plotting estimates used by C_{P-M} across levels of *P* and *S*. Although actual base rates and indicator skew exerted the most substantial influences on base-rate estimates, each method for estimating the base rate from MAXEIG curves was susceptible to biases whose direction and magnitude varied across conditions.

Discussion

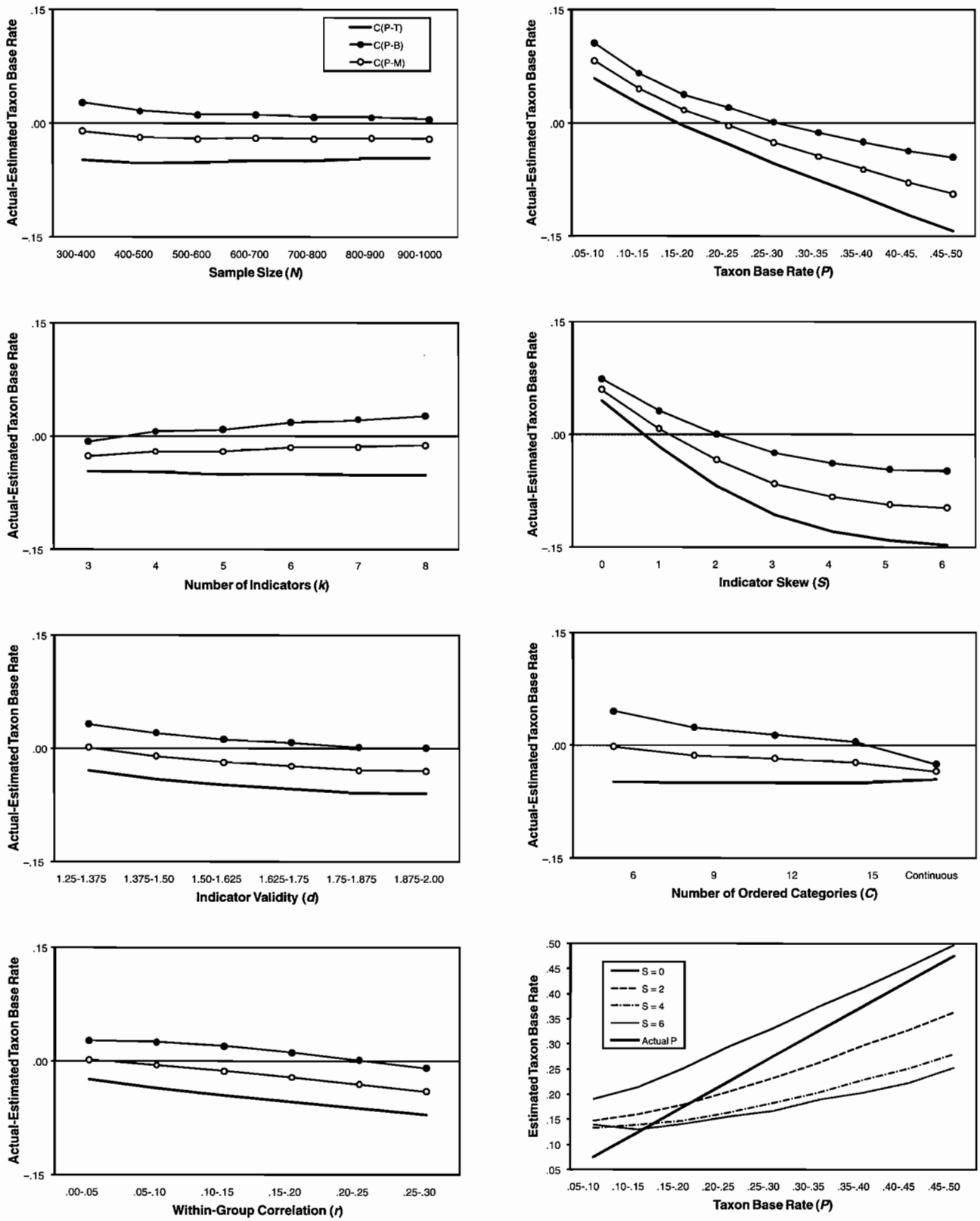
The central question that motivated this study was how best to classify cases using the results of taxometric analyses. When these results persuade a researcher that the latent variable under investigation is better fit by a taxonic than a dimensional structural model, the present findings suggest that the base-rate classification technique (C_P) will achieve greater accuracy than the Bayesian classification technique (C_{Bayes}). This general conclusion held not only across the full sample of 100,000 data sets, but also within every subsample across a wide range of data conditions. Not only does C_P yield better classification accuracy, but under most data conditions it also achieved a more even balance between sensitivity and specificity than C_{Bayes}.

To illustrate the consequences of misclassifying cases, suppose that one wants to assign individuals to MDD+ and MDD– groups to study their differences on demographic correlates of the disorder. The goal is to estimate the population mean difference (Δ), but misclassified cases in the sample of data at hand will reduce the observed mean difference (*d*), which in turn reduces the statistical power of a test between group means. To illustrate these effects, 10,000 samples of $n_1 = n_2 = 20$ were drawn from populations with normal distributions, equal variance, and $\Delta = 1.00$ (a large effect according to Cohen's, 1992, rules of thumb). Misclassifications were simulated by randomly swapping 0, 1, 2, 3, or 4 cases across groups, which corresponds to $HR = 1.00, 0.95, 0.90, 0.85$, and 0.80. Across these five conditions, *d* averaged 1.00, 0.90, 0.78, 0.67, and 0.56, and statistical power was .88, .78, .66, .53, and .39 for two-tailed *t* tests at $\alpha = .05$. The attenuation of *d* and loss of statistical power due to misclassified cases can be substantial, underscoring the importance of using the best available classification method.

Although it is clear that C_P classified cases more accurately than C_{Bayes} and that this can have important consequences, it is less clear which base-rate estimate should be used to implement the former method. The present study evaluated not only the use of the base-rate estimate obtained from a taxometric procedure—specifically, the mean of the estimates calculated from each MAXEIG curve—but also the use of a base rate estimated in a novel, two-step process: (a) use Bayes' theorem to assign cases to groups and (b) compute the proportion of cases assigned to the taxon as a new estimate of the taxon base rate. Although both of these estimates were biased under many data conditions—which is consistent with the results of an extensive study of taxometric base-rate estimation (Ruscio, 2007, Study 1)—the former tended to underestimate the actual taxon base rate by a larger amount than the latter tended to overestimate it; the same problem of underestimation usually was observed for an average of these two estimates. Clearly, more work remains to be done to improve the accuracy with which taxometric analyses estimate the taxon base rate, but the present findings suggest that researchers should consider calculating and reporting the proportion of cases assigned to the taxon using Bayes' theorem.

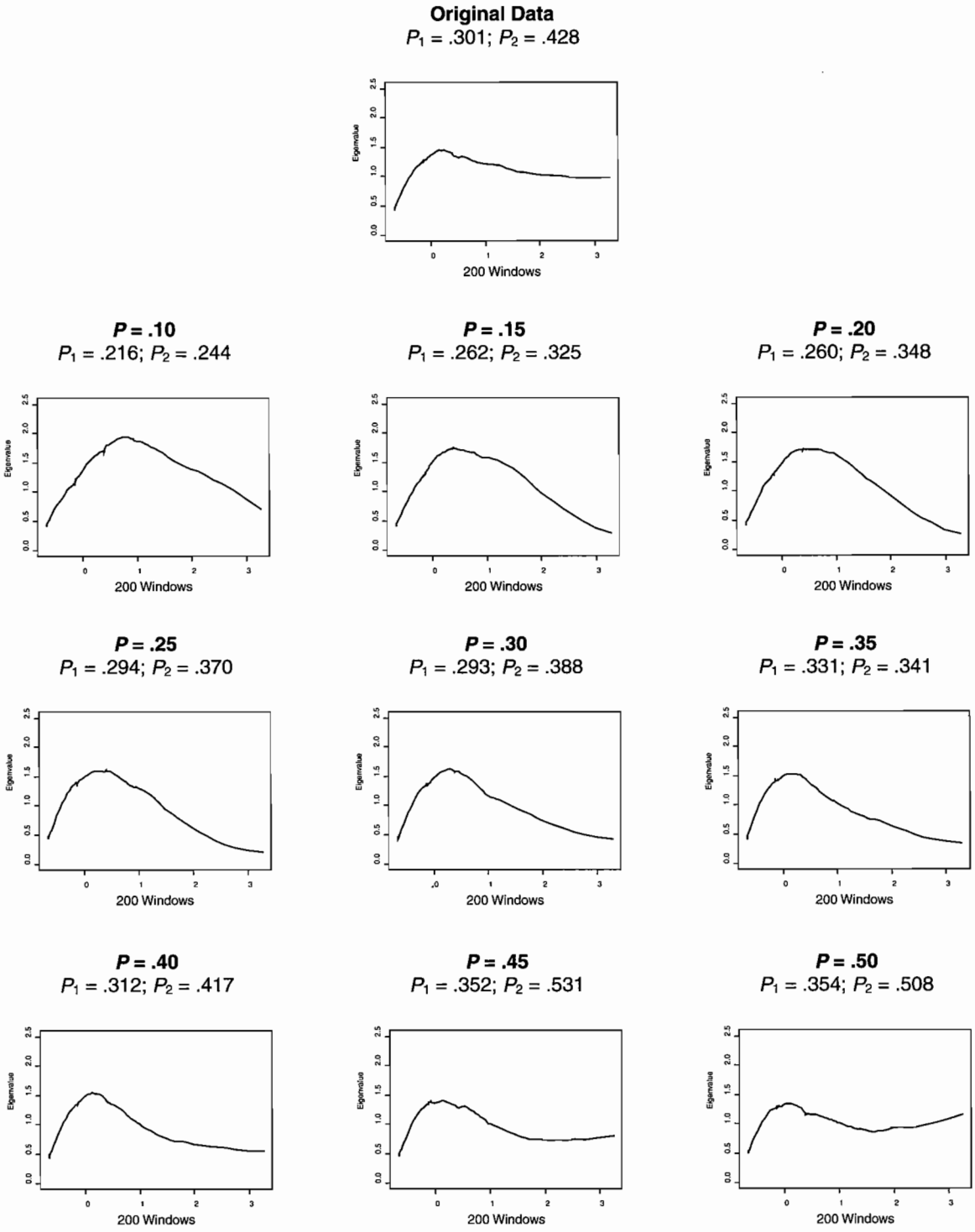
Because the C_{Bayes} method requires estimates of many parameters, it is not obvious how its accuracy might be improved. In contrast, the C_P method requires only an estimate of the taxon base rate. Any improvement in the estimation of taxon base rates would be expected to increase further the accuracy of C_P. One way to achieve this may be to utilize simulated comparison data in a new way. Ruscio et al. (2007) described how to generate taxonic and dimensional comparison data as an interpretive aid in taxometric analyses, and this approach could be adapted in an attempt to obtain more accurate estimates of the taxon base rate. For example, consider the results shown in the top cell of Figure 4, which contains an averaged curve from a MAXEIG analysis that

Figure 3
Discrepancies Between Actual and Estimated Taxon Base Rate Across Data Conditions



Note: Dotted lines at $y = .00$ represent accurate base-rate estimation. The graph in the lower right corner shows the estimated base rates (for the C_{P-M} method) across actual base rates and indictor skew.

Figure 4
Averaged MAXEIG Curves for an Original Taxonic Data set (top) and Nine Sets of Taxonic Comparison Data (Bottom Three Rows)



Note: P is the base rate used to generate each sample of taxonic comparison data, P_1 is the mean of the base-rate estimates calculated from the full panel of MAXEIG curves, and P_2 is the proportion of cases assigned to the taxon using Bayes' theorem. MAXEIG = MAXimum EIGenvalue.

produced base-rate estimates of $P_1 = .301$ and $P_2 = .428$ for a target data set; the former is the average of the estimates across MAXEIG curves, the latter the proportion of cases assigned to the taxon using Bayes' theorem. A researcher would not know the actual taxon base rate ($P = .40$ in this example), and the results of the present study show that either estimate (or both) can be biased. How might analyses of taxonic comparison data help to obtain an accurate estimate?

Using estimates spanning a broad range of values—in this case, from $P = .10$ to .50 in increments of .05—multiple taxonic comparison data sets were generated and submitted to MAXEIG analysis. This yielded the results shown in the bottom three rows of Figure 4. By comparing the MAXEIG results for the original data set to those for each comparison data set, one can determine which comparison data set best reproduces the original. In this case, it appears that the results highlighted in the lower left corner, for the taxonic comparison data generated using $P = .40$, are most similar to those for the original data. Not only are the base-rate estimates the most similar, but so is the relative flatness of the MAXEIG curve at the right end. All taxonic comparison data generated using $P < .40$ yielded lower base-rate estimates and curves that sloped downward at the right, whereas comparison data generated using $P > .40$ yielded higher base-rate estimates and curves that sloped upward at the right. Thus, this approach provided a base-rate estimate of .40—which in this case is the correct value.

Of course, a single demonstration such as this may represent nothing more than a fortunate coincidence, and readers may disagree with the author's conclusion that the results for $P = .40$ do in fact match those for the original data most closely. This is not presented as evidence to support this approach, only to suggest that it may be useful and merits further study. This approach could yield more accurate base-rate estimates than currently available methods because each comparison data set reproduces the distributional and correlation characteristics of the original data set, thereby holding constant a number of factors that have been shown to bias other estimates. For simplicity, this demonstration involved only one set of taxonic comparison data at each base rate, and these base rates varied rather coarsely, in increments of .05. For a more rigorous test with greater fidelity, multiple sets of taxonic comparison data should be generated at each base rate and a finer gradation of base rates should be used.

When a taxonic structural model appears to provide better fit than a dimensional model, C_P provides a straightforward and surprisingly effective way to assign cases to groups. Because it requires only an estimate of the base rate of taxon membership, this approach also can be implemented for dimensional data when practical considerations require classification (e.g., provision of a resource-intensive treatment to some, but not all, clients, or reporting demographic characteristics of diagnosable vs. nondiagnosable individuals for administrative or research purposes). In this situation, applying Bayes' theorem using the results of a taxometric analysis poses significant difficulties. For example, estimating each variable's true and false positive rates depends on the identification of an optimal threshold, ordinarily suggested by a peak in a taxometric curve. When data are not taxonic, however, there may be no peak to guide the selection of a threshold.

Taking advantage of C_P does not depend on obtaining any particular pattern of taxometric results. All that is required is an estimate of the taxon base rate. When this cannot (or should not) be calculated from taxometric results (e.g., when they are more consistent with dimensional than taxonic structure), it can be obtained in other ways. For example, one can examine the frequency with which individuals meet pertinent diagnostic criteria, seek treatment, or qualify to receive certain benefits or services. Once a base rate is estimated, using it to place a threshold along a total score constructed from available data is a simple and familiar operation. The present study suggests that C_P achieves an impressive accuracy level when a taxonic model appears to fit the data better than a dimensional model, but it can be applied even when this is not true yet the circumstances nonetheless call for the assignment of cases to groups. Given the relative ease with which base rates can be estimated—whether from the results of a taxometric analysis, well-supported theory, previous research findings, or other sources of information—C_P appears to be a simple, broadly applicable, and highly effective method for assigning cases to groups using available data.

Appendix
Program to Generate Data for Each Group

Readers interested in running these programs are encouraged to contact the author for an electronic copy.

```
GenData <- function(N, k, r, S, C, Multiplier = 1)
{
        Freq.Dist <- matrix(nrow = N, ncol = k)
        for (i in 1:k)
        {
                if (S == 0) x <- rnorm(N)
                else
                {
                        if (S == 1) b <- 1.1038035
                        if (S == 2) b <- 1.3553013
                        if (S == 3) b <- 1.668685
                        if (S == 4) b <- 2
                        if (S == 5) b <- 2.332112
                        if (S == 6) b <- 2.6587115
                        x <- Lognorm(N, 1, b)
                }
                if (C == 0) Freq.Dist[,i] <- x
                else
                {
                        Prop.Lo <- 1
                        Iter <- 0
                        while (Prop.Lo > .95)
                        {
                                Trimmed.x <- sort(x)[1:round(((10 - Iter)/10) * N)]
                                Min.x <- min(Trimmed.x)
                                Max.x <- max(Trimmed.x)
                                Cuts <- seq(from = Min.x, to = Max.x, length = (C + 1))
                                for (j in 1:N)
                                        for (l in 1:C)
                                                if (x[j] >= Cuts[l]) Freq.Dist[j, i] <- l
                                Prop.Lo <- sum((Freq.Dist[,i] < 2) * 1) / N
                                Iter <- Iter + 1
                        }
                }
                Freq.Dist[,i] <- sort(Freq.Dist[,i])
        }
        Target.Corr <- diag(k)
        Target.Corr <- Target.Corr + r - (diag(k) * r)
        Desired.Corr <- Target.Corr
        Shared.Comp <- rnorm(N, mean = 0, sd = 1)
        Unique.Comp <- matrix(rnorm(N * k, mean = 0, sd = 1), nrow = N, ncol = k)
        y <- matrix(0, nrow = N, ncol = k)
        Iter <- 0
        Best.RMSR <- 1
        j <- 0
        while (j < 5)
        {
                Iter <- Iter + 1
                Shared.Load <- Factor.Analysis(Desired.Corr, Corr.Matrix = T, N.Factors = 1)$loadings
                Shared.Load[Shared.Load > 1] <- 1
                Shared.Load[Shared.Load < -1] <- -1
                Unique.Load <- sqrt(1 - Shared.Load ^ 2)
                for (i in 1:k)
                        y[,i] <- Shared.Load[i] * Shared.Comp + Unique.Load[i] * Unique.Comp[,i]
                for (i in 1:k)
                {
                        y <- y[sort.list(y[,i]),]
                        y[,i] <- Freq.Dist[,i]
                }
                Reproduced.Corr <- cor(y)
                Residual.Corr <- Target.Corr - Reproduced.Corr
                RMSR <- sqrt(sum(Residual.Corr[lower.tri(Residual.Corr)]^2) /
                                length(Residual.Corr[lower.tri(Residual.Corr)]))
                if (RMSR < Best.RMSR)
                {
                        Best.RMSR <- RMSR
                        Best.Corr <- Desired.Corr
                        Best.Res <- Residual.Corr
                        Desired.Corr <- Desired.Corr + Multiplier * Residual.Corr
                        j <- 0
                }
```

(continued)

Appendix (continued)

```
                else
                {
                        j <- j + 1
                        Mult <- Multiplier / (2 ^ j)
                        Desired.Corr <- Best.Corr + Mult * Best.Res
                }
        }
        Shared.Load <- Factor.Analysis(Best.Corr, Corr.Matrix = T, N.Factors = 1)$loadings
        Shared.Load[Shared.Load > 1] <- 1
        Shared.Load[Shared.Load < -1] <- -1
        Unique.Load <- sqrt(1 - Shared.Load ^ 2)
        for (i in 1:k)
                y[,i] <- Shared.Load[i] * Shared.Comp + Unique.Load[i] * Unique.Comp[,i]
        for (i in 1:k)
        {
                y <- y[sort.list(y[,i]),]
                y[,i] <- Freq.Dist[,i]
        }
        y <- apply(y, 2, scale)
        return(y)
}


Lognorm <- function(N, a, b, St = T)
{
        x <- exp(log(a) + sqrt(log(b)) * rnorm(N))
        if (St)
        {
                M <- a * exp(log(b) / 2)
                SD <- sqrt(a ^ 2 * b * (b - 1))
                x <- (x - M) / SD
        }
        return(x)
}


Factor.Analysis <- function(Data, Corr.Matrix = F, Max.Iter = 50, N.Factors = 0)
{
        Data <- as.matrix(Data)
        I <- dim(Data)[2]
        if (N.Factors == 0) N.Factors <- I
        if (!Corr.Matrix) Cor.Matrix <- cor(Data)
                else Cor.Matrix <- Data
        Criterion <- .001
        Old.H2 <- rep(99, I)
        Change <- 1
        Iter <- 0
        Factor.Loadings <- matrix(nrow = I, ncol = N.Factors)
        H2 <- vector("numeric", I)
        while ((Change >= Criterion) & (Iter < Max.Iter))
        {
                Iter <- Iter + 1
                Eig <- eigen(Cor.Matrix)
                L <- sqrt(Eig$values[1:N.Factors])
                for (i in 1:N.Factors)
                        Factor.Loadings[,i] <- Eig$vectors[,i] * L[i]
                for (i in 1:I)
                        H2[i] <- sum(Factor.Loadings[i,] * Factor.Loadings[i,])
                Change <- max(abs(Old.H2 - H2))
                Old.h2 <- H2
                diag(Cor.Matrix) <- H2
        }
        if (N.Factors == I)
                for (i in 1:I)
                        if (Eig$values[i] > 1) N.Factors <- i
        return(list(loadings = Factor.Loadings[,1:N.Factors], factors = N.Factors))
}
```

Notes

1. For taxon membership to differ even when the sizes of the taxa yielded by C_{Bayes} and C_P are identical, two conditions must be met. First, there must be discrepancies in the rank-ordering of cases' total scores and Bayesian probabilities of taxon membership. This usually will be the case, because Bayes' theorem treats any two above-threshold responses—or any two below-threshold responses—to the same item as functionally equivalent. For example, suppose that the optimal threshold for each of four Major Depressive Disorder (MDD) symptom rating scales is five and that three individuals' response patterns are as follows: {4, 4, 4, 4}, {3, 3, 3, 6}, and {1, 1, 6, 6}. Total scores descend (16, 15, 14) for these three cases, yet their probabilities of taxon membership must ascend because the first case has no above-threshold responses, the second case has one (Item 4), and the third case has the same as the second (Item 4) plus another (Item 3). Thus, there are discrepancies in the ranks. The second condition is that these discrepancies must have consequences for taxon membership—they cannot amount to shuffling ranks only within groups. Suppose that each method assigned just one of the three cases listed above to the taxon: Bayes' theorem would assign only the third case (with two above-threshold scores), whereas only the first case would be assigned to the taxon on the basis of having the highest total score. Even if each method assigned two cases to the taxon, group membership would differ across methods: For C_{Bayes}, the second and third cases would be assigned to the taxon; for C_P, the first and second cases would be assigned to the taxon. In this example, rank discrepancies have consequences for taxon membership.

2. To estimate the true and false positive rates achieved by the optimal thresholds for each of k variables, a series of MAXimum COVariance (MAXCOV) or MAXimum EIGenvalue (MAXEIG) analyses must allow each variable to serve as the input indicator. For MAXEIG, all remaining variables serve simultaneously as output indicators, yielding k analyses. For MAXCOV, pairs of remaining variables serve as output indicators, yielding a total of $k(k-1)(k-2)/2$ analyses. For $k > 3$, MAXCOV requires more analyses than MAXEIG (e.g., at $k = 8$ MAXEIG requires only 8 analyses but MAXCOV requires 168).

References

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

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Annals of Internal Medicine, 138*, 40-45.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

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

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl (Eds.), *Psychodiagnosis: Selected papers* (pp. 200-224). Minneapolis: University of Minnesota Press.

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

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

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: John Wiley.

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

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

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

Ruscio, J. (2007). Taxometric analysis: An empirically-grounded approach to implementing the method. *Criminal Justice and Behavior, 24*, 1588-1622.

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

Ruscio, J., & Kaczetow, W. (in press). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*.

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. (2002). A structure-based approach to psychological assessment: Matching measurement models to latent structure. *Assessment, 9*, 4-16.

Ruscio, J., & Ruscio, A. M. (2004a). A conceptual and methodological checklist for conducting a taxometric investigation. *Behavior Therapy, 35*, 403-447.

Ruscio, J., & Ruscio, A. M. (2004b). A nontechnical introduction to the taxometric method. *Understanding Statistics, 3*, 151-193.

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.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*, 209-219.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

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

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.