Holden, J. E., & Kelley, K. (2008, March). Effects of misclassified data on two methods of classification analysis: A Monte Carlo simulation study. A paper presented at the annual meeting of the American Educational Research Association, New York, NY.

# The Effects of Initially Misclassified Data on the Effectiveness of Discriminant Function Analysis and Finite Mixture Modeling

Jocelyn E. Holden

Indiana University Bloomington

Ken Kelley

Indiana University Bloomington

Authors Note:  The authors would like to thank Michael P. Trelinski for all his computer programming insight.

## Abstract

Classification procedures are common and useful in behavioral, educational, and social research. Supervised classification techniques like discriminant function analysis (DFA) assume training data is perfectly classified when estimating parameters and classification. In contrast, unsupervised classification techniques like finite mixture models (FMM) do not require knowledge of group status in order to estimate parameters or classify. The purpose of this study is to investigate the impact of misclassification errors (both randomly distributed errors and errors weighted toward the distribution overlap) on the classification accuracy of DFA and FMM. Analytic and Monte Carlo results are provided for a variety of misclassification scenarios to investigate the performance of the two methods. DFA recovered the highest overall percentages of correctly classified data, whereas FMM captured higher percentages of the smaller group when group sizes are unequal.

Classification of individuals into non-overlapping groups is regularly used in the behavioral, educational and social sciences as well as in many other fields. Classification is a fundamental part of organization in most fields (Keogh, 2005). Zigler and Phillips (1961) argue that there are three criteria important in determining the appropriateness of a classification scheme: homogeneity (the similarity of individuals in categories), reliability (consistency or agreement among who should be included in a category), and validity (how well category membership informs us about their characteristics). All three of these criteria will be at risk when errors in statistical classification are made.

Classification can refer to observed groups (e.g. sex, grade level, employment status, etc.) or latent classes (e.g. learning disabled, depressed, or alcoholic).[1] When class is latent, or unobservable, misclassification errors are almost unavoidable since it is generally not possible to know an individual's true class. Even when class is directly observable, classification mistakes can be made. Misclassification errors can arise from use of crude or imprecise classification methods due to budgetary, time, personnel constraints, or to practical constraints on data collection procedures (Bross, 1954; Katz & McSweeney, 1979).

Misclassification can also arise from the nature of statistical classification methods. Aside from external problems with untrustworthy data collection, statistical classification itself is essentially never 100% accurate in practice. Statistical classification methods can only be as good as the predictors chosen for the study. To put a further cap on classification accuracy, evidence from many simulation studies on classification analysis, perhaps not surprisingly, reveals statistical classification methods to have varying accuracy under a variety of different situations. For example, it has been found in many studies that the ratio of group sizes makes a large difference in the ability of classification analyses to correctly classify cases (Finch &

Schneider, 2006). A large degree of overlap between samples (Blashfield, 1976; Harrell & Lee, 1985) and lack of sphericity have also been found to lead to inaccurate classification (Blashfield, 1976; deCraen et al., 2006). Inaccuracy of classification due to lack of sphericity may be somewhat alleviated if groups or classes to be recovered are more unequal in size (deCraen et al., 2006). Different forms of error perturbation (Baker, 1979; Breckenridge, 2000) have been shown to reduce classification accuracy for cluster analytic techniques. Furthermore it has been found that for discriminant function analysis both outliers and inliers in the training data set can pose problems not only for classification accuracy (Kuiper & Fisher, 1975; Van Ness & Yang, 1998), but can also lead to serious underestimates of the accuracy of the analysis (Edelbrock, 1979).

Interestingly, some authors have found that, greater numbers of true clusters existing in the data can lower the misclassification rate (Kuiper & Fisher, 1975; Milligan, Soon & Sokol, 1983), whereas others have found that greater numbers of clusters lead to higher rates of misclassification (Breckenridge, 2000). Increased number of variables used in prediction (Breckenridge, 2000; Lubke & Muthen, 2007), goodness of model fit (Breckenridge, 2000), accuracy of prior probabilities (Lei & Koehly, 2003), and standardization of data (Edelbrock, 1979) have also been shown to lead to less error in classification.

Classification analysis methods are ubiquitous in nearly all areas of research. Methods such as discriminant function analysis or logistic regression are often used to help researchers better understand the characteristics of individuals belonging to particular groups, as well as to discover the best variables, and methods for differentiating between groups. Many important decisions are made based on how cases or individuals are classified (e.g. does a student pass or fail a course, does an individual receive treatment for depression, etc.). Thus it is very important to keep the limitations of classification analyses in mind when interpreting results.

The discussion thus far of misclassification errors was restricted to situations where initial knowledge of correct classification is either not necessary (e.g. in the context of cluster analysis and mixture models) or in the case of models which require initial knowledge to be perfect (e.g. in the context of discriminant function analysis). However, what happens when the data classification method best suited to the research purpose relies on initial knowledge of correct classification, but the available data has misclassification errors? Such errors could be due to data entry mistakes, low reliability, or perhaps most likely, inadequate or ambiguous identification schemes (e.g., behavior disorders, learning disabilities, or concept mastery)? How much does initial misclassification of training data impact the ability of classification analysis schemes to accurately recover groups?

When scales or tests of an attribute are involved, researchers have proposed methods for estimating the number of individuals who may be misclassified by the instrument (Rudner, 2001; 2005). According to Bross (1954), at that time opinions about the effects of misclassified data ranged from refusing to work with data containing possible misclassifications because the data are flawed, to believing the effects of misclassified data will cancel each other out and thus not pose a problem. Unfortunately since Bross's publication there have been very few studies on this specific topic. In general, the few existing studies have focused on the effects of initially misclassified data on chi-square analyses, and the available literature appears to advocate a more intermediate position. For chi-square, it has been found that misclassification of data categories does not impact the validity of the test of significance, although it may reduce the power of the test (Bross, 1954; Assakul & Proctor, 1967; Katz & McSweeney, 1979). However, chi-square is just one of many different forms of classification analysis. Statistical proofs by Lachenbruch (1966) indicate that initial misclassification of data in discriminant function analysis may impact classification accuracy but it does so only a small amount. The conditions under which Lachenbruch (1966) tested discriminant function analysis were quite restrictive (assumed misclassification to be symmetric and at random) and did not take various mediating variables into account. Thus, as it stands, very little is known about the magnitude of the impact of misclassified training data on the accuracy of classification analysis in conditions of interest to those in the behavioral, educational and social sciences.

The purpose of the present research is to discern the effect initially misclassifying data has on the effectiveness of the two group case of discriminant function analysis (DFA) and finite mixture modeling (FMM) under various data and distributional characteristics. This comparison of methods was chosen because DFA requires training data (supervised classification) while FMM classification is based on a probabilistic model where group membership is unknown and thus does not require prior group classification (unsupervised classification). Although there are several other methods of supervised classification (e.g. logistic regression or chi-square) and unsupervised classification (e.g. cluster analytic models) to keep things simple, this study was limited to one type of each (DFA and FMM) due to the popularity of each. Because the DFA is based on more information than FMM, we make the following hypotheses:

*Hypothesis 1: When training data is perfectly classified, discriminant function analysis will provide more accurate classifications than finite mixture models.*

*Hypothesis 2: As misclassified data is increasingly introduced into the samples, the ability of discriminant function analysis to provide accurate classifications will continue to decrease while the finite mixture model accuracy will remain unchanged.*

An examination of the mathematics involved in DFA and FMM also dictate when effects of misclassification should be observed. Recall that the procedure for discriminant function analysis involves choosing the linear combination of variables that maximizes the multivariate distance between groups, termed the discriminant function, and then based on this discriminant function a decision rule is constructed which classifies each individual in the group to which their specific discriminant function score is most similar. Let **a** be a vector of coefficients, **S** be the unbiased pooled estimate of the population covariance matrix (which assumes homogeneity of population covariance matrices: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the common population covariance matrix), $\bar{\mathbf{x}}_i$ be the mean vector of length *p*, where *p* is the number of variables for the cases in group *i*, and $D(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)$ be the multivariate distance between $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ which is defined as

(1) $$D(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \frac{\left|\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right|}{(\mathbf{a}'\mathbf{S}\mathbf{a})^{1/2}} .$$

When $D(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)$ is maximized by a particular vector **a**, the resulting **a** vector becomes the vector of discriminant function coefficients. Conceptually, the multivariate distance between $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ is the maximum of the univariate distances between $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ (Flury, 1997). It becomes obvious that the discriminant function equation used in classification is dependent on the means of the scores in each group, the group centroids, and the common covariance matrix. The group centroids are calculated from the raw data as

(2) $$\bar{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_i} x_{ij} ,$$

where *j* = 1,2 for the group in which the case is classified and $n_j$ is the number of entities in group *j*. Thus, theoretically speaking, accuracy in classification is contingent on Equation 2 yielding accurate estimates of the mean vectors for each group.[2] When group status (*j*) is incorrect for certain individuals, centroids yielded from Equation 2 are less likely to represent the true population values. Following from this, any classifications made from these biased centroids will contain some degree of classification inaccuracy. Thus, based on the discriminant function analysis equations, the accuracy of DFA solutions should decrease as the amount of misclassified data is increased as such misclassification will tend to bias the mean vectors.

Unlike discriminant function analysis, it becomes obvious looking at the definitional equations for finite mixture modeling, that initial misclassification of group status will have no effect on the results. The sample composite distribution function based on the finite mixture model has the form

(3)
$$f(x) = \sum_{m=1}^{M} \hat{\pi}_m \varphi(x; \bar{\mathbf{x}}_{\mathbf{m}}, \mathbf{S}_{\mathbf{m}})$$

where $\pi_m$ are the sample mixing proportions ( $\sum_{m}^{M} \hat{\pi}_n = 1$ ), and φ represents the normal distribution function with sample mean $\bar{\mathbf{x}}_{\mathbf{m}}$ and covariance matrix $\mathbf{S}_{\mathbf{m}}$**,** also called component distributions because they are the distributions which comprise the finite mixture distribution (i.e the composite). The basis for classification in FMM is the posterior probabilities. The posterior probability is the probability that an entity belongs to distribution A or distribution B of the fitted model. To estimate the posterior probabilities of group membership, the following equation is used:

(4)
$$\hat{r}_{im} = \frac{\hat{\pi}_m \varphi(x_i; \hat{\mathbf{x}}_{\mathbf{m}}, \hat{\mathbf{S}}_{\mathbf{m}})}{\sum_{m=1}^{M} \hat{\pi}_m \varphi(x_i; \hat{\mathbf{x}}_{\mathbf{k}}, \hat{\mathbf{S}}_{\mathbf{k}})}$$

where $\hat{r}_{im}$ is the estimated posterior probability of $x_i$ belonging to component distribution m, and $\hat{\pi}_m, \hat{\boldsymbol{\mu}}_{\mathbf{m}}$, and $\hat{\mathbf{S}}_{\mathbf{m}}$ are the estimated mixing proportions, mean vectors and covariance matrices (for FMM it is not necessary to assume equal covariance matricies) for the component distributions, respectively (Hastie, Tibshirani & Friedman, 2001). In finite mixture modeling, cases are classified as belonging to the distribution to which they have the highest posterior probability.

For two groups, this is the same as setting a cut point (c) where the posterior probabilities at y = c are 0.5 (Flury, 1997). As can be seen from Equation 4, classification in a finite mixture model is a function of estimated mixing proportions, means and covariance matrices for the component distributions. Notice that the known or estimated initial group status, if one is available, does not enter the equation. Thus in the case of the FMM, no initial group status is needed, nor is the accuracy of initial classification an issue because this information is not used by the model.

From a conceptual and mathematical perspective, our hypotheses are justified. Based on the research results summarized earlier, it is also reasonable to hypothesize that different characteristics of the data may interact with the ability of DFA and FMM to classify accurately when training data is misclassified above and beyond any "main effects" that may exist. As demonstrated, many studies have reported circumstances and data characteristics which affect classification accuracy when initial data classification is perfect. Such characteristics include sample size, group size ratio, distance between group means, and variance of distributions. Based on these results, a third hypothesis is justified:

*Hypothesis 3: Manipulation of data and distribution characteristics, such as sample size, effect size, and sample size ratio will lead to differences in classification accuracy. In the case of discriminant function analysis these characteristics may interact with initially misclassified data proportions to produce poorer misclassification.*

Because misclassification problems can take different forms, it was thought important to investigate the effects of two different types of misclassification. The two types of misclassification examined were random and weighted (See Figure 1). *Random misclassification* refers to the situation where any individual or case has the same probability of being misclassified as any other data point regardless of relative position in the distribution. This is

----Figure 1 About Here----

thought to be analogous to situations where misclassification enters datasets through data entry mistakes or human error. *Weighted misclassification* refers to the situation where, depending on the relative position in the distribution, data points have differing probabilities of being misclassified. In particular, points closer to the overlap of the distributions would be more likely to be misclassified than points lying on the outer tails of the distributions. This situation is thought to be most analogous to classification in diagnostic categories such as learning disabilities, depression or alcoholism. For such scenarios, misclassification will not tend to be random, but rather borderline cases will be misclassified at a higher rate.

## STUDY 1

In study 1, data were generated to simulate the situation where data is initially misclassified at random. Once generated, data were analyzed with both discriminant function analysis and finite mixture modeling in an attempt to recover the true groups.

### *Methods*

#### Data Generation

Generation and analysis of misclassified data was accomplished using the R statistical software package (R Development Core Team, 2007). Data were generated to meet the specific data and distribution criteria described in Table 1. Each condition is completely crossed with all other conditions for a total of 216 (4 x 3 x 3 x 6) conditions. The standardized mean difference was used as a measure of effect size. The particular values of effect size were chosen to coincide with Cohen's (1988) guidelines for a small (.2,), medium (.5), and large (.8) effect, as well as a very small (.1, 1/2 the size of small) and very large (1.6, twice the size of large).

-----Table 1 About Here----

For the purposes of this paper, data generation was limited to symmetric misclassification. In other words, equal percentages from each distribution will be misclassified. For example, 10% misclassification implies 10% of group A misclassified as B and 10% of B misclassified as A. To simplify interpretation, data were also limited to one predictor variable. A raw percentage (number of cases correctly classified divided by the total number of cases) is used as a measure of the amount of cases correctly classified.

To achieve data misclassified symmetrically at random, the following procedure was used: First, the R function `rnorm()` was used to create 2 classes of data with specified means and standard deviations. The distribution with the smaller mean was labeled "distribution A", the distribution with the larger mean was labeled "distribution B". In Study 1, misclassification was at random making every point is just as likely to be misclassified as every other point. For each point, a random number between 0 and 1 (using the `runif()` function) was generated. If the number was smaller than the desired misclassification percentage (e.g. 0.1) the point would be relabeled as belonging to the other distribution. In other words, if we want 10% of the data to be correct, points with random numbers less than .1 would be misclassified thus ensuring a random 10% of the cases would be misclassified on every iteration.

Analyses

After data were misclassified, a finite mixture model and a discriminant function analysis were performed on the misclassified data. The finite mixture model was run using the `Mclust()` function located in the `mclust` R package (Fraley & Raferty, 2002), with two groups specified. For each iteration, the FMM classes were labeled such that the labeling scheme, which achieved the highest percentage correct for each group was chosen. The discriminant function analysis was performed using the `lda()` function located in the `MASS` R package (Venables & Ripley, 2002). For both DFA and FMM, the default settings for the analysis were used. For FMM variances were assumed unequal, and for DFA variances were assumed equal and prior probabilities were taken from the data. Such a condition was used because FMM generally assume heterogeneous variances whereas DFA generally assumes homogeneous variances. Thus, to be consistent with what tends to appear in the literature, different error structures were used in when evaluating the effectiveness of the procedures. At the completion of each analysis, the classification results were compared to the true data (not the misclassified data) and a percent correct was calculated for each. Also calculated are the percent of group A cases misclassified as group B, and the percent of group B cases misclassified as group A. Analysis of these percents allows investigation into whether, and under what conditions, the methods may be biased toward misclassification in one direction. 10,000 replications of this procedure were executed for each of the 216 simulation conditions. R code is available from authors upon request.

*Results*

The percents of correctly classified cases for each classification method are displayed in Tables 2 and 3 (although three sample sizes were tested, for space considerations results are only displayed for the largest and smallest samples. Results for the intermediate sample size condition were completely consistent with results from small and large sample sizes. Readers can request full set of results from authors). In line with previous research (Finch & Schneider, 2006) and consistent with the discriminant function, it can be seen that as effect size and sample size increase, the ability of both classification analyses to correctly classify data increases.

Percent correct for DFA ranged from .500 - .788 for the 50:50 condition, .750 -.835 for the 25:75 condition, and .900 - .914 for the 10:90 condition. Percent correct for FMM ranged from .513 - .785 for the 50:50 condition, .542 - .819 for the 25:75 condition, and .565 - .835 for the 10:90 condition. DFA outperformed FMM in the majority of conditions overall. However, FMM showed higher levels of classification accuracy in the 50:50 condition at the lower effect sizes.

Consistent with the findings of Breckenridge (2000), deCraen et al. (2006), and Finch and Schneider (2006), sample size ratio has a substantial impact on the effectiveness of both FMM and DFA. In particular, as group size becomes more and more discrepant, the ability of both FMM and DFA to classify correctly increases. For FMM, the increase in classification accuracy is only marginal, but for DFA the increase is quite large. For FMM, there is a possible interaction between sample size and effect size. For example compare $\delta = 0$ and $\delta = 1.6$ for $N = 100$. For the smaller effect size we observe 54% correct classification for the 50:50 sample size ratio, and 66% correct classification for the 10:90 sample size ratio: a difference of 12 raw percentage points. For the larger effect size we observe 75% correct for the 50:50 sample size ratio and 76% correct classification for the 10:90 sample size ratio: a difference of 1 raw percentage point. When the sample size is increased, however, the increase in classification accuracy due to sample size ratio increases for the larger effect sizes and decreases for the smaller effect sizes. Compare the $\delta$ = 0 and $\delta$ = 1.6, N = 1000. For the smallest effect size we observe 51% classification accuracy for the 50:50 sample size ratio and 57% correct for the 10:90 sample size ratio: a difference of 6 raw percentage points. However at the largest effect

----Tables 2, 3, 4 and 5 about here----

size we observe 79% correct for the 50:50 sample size ratio and 83% correct for the 10:90 sample size ratio: a difference of 4 raw percentage points. Thus for the FMM, as sample size increases, the effect of sample size ratio on classification accuracy changes depending on effect size. Introduction of misclassified cases did not impact the finite mixture model classification in any way as initial classification is not part of the model. There seemed to be no interaction with sample size ratio or effect size.

As expected, no effect of misclassified data on the misclassification direction of FMM was observed. However, when looking at misclassification for DFA we see marginal effects of misclassified data. As the percent of misclassified data increases, the percent of the smaller group misclassified as the larger group decreases, and the percent of the larger group misclassified as the smaller group increases.

Arguably the most interesting result is that FMM and DFA tend to misclassify in different directions, which has not yet been reported in the literature. When group sizes are equal, approximately equal numbers of cases are misclassified in either direction for both DFA and FMM. However, when group sizes are unequal, FMM misclassifies in favor of the smaller group whereas DFA misclassifies in favor of the larger group. In other words, as sample size ratio becomes more discrepant, FMM misclassify fewer cases from the smaller group, and DFA misclassifies fewer cases from the larger group. As sample size and effect size are increased, these numbers are decreased even further. To illustrate, when sample size is 100 and the sample size ratio is very discrepant (10:90), FMM misclassify between 4% and 7% of the smaller group while DFA misclassifies between 8% and 10% (depending on effect size and percent misclassification). On the other hand, DFA misclassifies less than 2% of the larger group while FMM misclassifies between 19% and 27%. Suppose the smaller group as a target population to be identified (for example, learning disabled or at risk students, alcoholics, or individuals with cancer), DFA can be said to miss a higher percentage of the target population, while FMM can be said to have a higher rate of false positives (see Table 6). To summarize, these results indicate that although DFA recovers higher percentages of correctly classified cases, FMM does a better job of recovering the smaller group.

----Table 6 About Here---

STUDY 2

Study 2 followed the same procedure as study 1 except data were generated to simulate weighted misclassification errors. In many cases, especially in contexts where a cut-point is used to place individuals or cases into different categories, misclassifications occur with a higher probability for certain individuals than for others. Very few instruments can reliably distinguish between cases with adjacent scores (Dwyer, 1996). Thus in these contexts cases nearer to the cut-point are more likely to be misclassified than cases lying further away (Lathrop, 1986; Dwyer, 1996). To serve as an analog to these situations, study 2 simulated data such that cases with a low probability of belonging to their parent distribution were more likely to be misclassified than cases with a high probability of belonging to their parent distribution.

*Methods*

Data Generation

As in study 1, data generation and analysis were conducted using the R software package and the raw percentage used as a measure of classification accuracy. Data were generated to the same conditions as in study 1 except data misclassification was weighted, not random. Cases with the lowest probability of belonging to their parent distribution were more likely to be misclassified than cases with a higher probability of belonging to their parent distribution. To achieve this type of misclassification, a procedure similar to study 1 was used. Instead of generating a random number between 0 and 1 for each case and comparing to the percent misclassified overall, the random number was compared to a scaled cumulative probability of the point (thus weighing unlikely cases more likely to be misclassified) so that we could control the proportion of misclassified data. When this procedure is used, without a scalar approximately 50% of the cases will be misclassified every time. By multiplying by an appropriate scalar, $k$, the percent of data misclassified could either be raised or lowered by lowering all of the probabilities by the same amount making it either more or less likely for data to be misclassified. An optimization program was then written to determine the values of $k$ that were necessary to achieve 0%, 10%, 20% and 30% weighted misclassification.

Analyses

As in study 1, a finite mixture model and a discriminant function analysis were run and the result was compared to the known classification. Again, the percent of group A cases misclassified as group B, and the percent of group B cases misclassified as group A were calculated.

Results

Results from the largest and smallest sample sizes from the weighted condition are presented in Tables 7 – 10. In general, the results using weighted misclassified data mirrored the results from the randomly misclassified data in study 1. As in the random condition, classification accuracy increased slightly with the increase in sample size, and increased dramatically with increase in effect size and sample size ratio. As before, the initial misclassification does not affect the FMM solutions. Also, as expected, we see a decrease in DFA classification accuracy as the percent of misclassified training data in the sample is increased.

Percent correct for DFA ranged from .499 - .788 for the 50:50 condition, .604 - .819 for the 25:75 condition, and .750 - .915 for the 10:90 condition. In comparison, percent correct for FMM ranged from .535 - .788 for the 50:50 condition, .603 - .819 for the 25:75 condition and 566 - .834 for the 10:90 condition. Overall, DFA resulted in higher percentages of correctly classified data than did FMM. However, for the 50:50 condition at the lower effect sizes, FMM provided more accurate classification than did DFA.

Again replicating the results of study 1, the same pattern regarding the increase in classification due to sample size ratio emerges. For DFA the increase in accuracy due to sample size ratio is quite large, whereas, the increase is only marginal and appears to be reduced as effect size increases for FMM. It should be noted that, we see exactly the same patterns for FMM in study 1 and study 2 because FMM do not take training data into account. It does not matter how the data were misclassified, the results will be identical. Thus the FMM results of study 1 and study 2 are exact replications of each other.

When looking at the direction of misclassification (shown in Tables 9 and 10) a pattern

----Tables 7, 8, 9 and 10 about here---

similar to that in study 1 emerges. Recall that in study 1, when data were misclassified at random it was observed that FMM and DFA misclassify cases in opposite directions: FMM tends to misclassify in favor of the smaller group while DFA tends to misclassify in favor of the larger group. When data were misclassified systematically, in general we see the same pattern. However, in the 70% correct training data condition and the highest effect sizes of the 80% correct condition the pattern changes. The FMM pattern stays the same, but for DFA, the direction of misclassification reverses: DFA begins to misclassify in favor of the larger group instead of the smaller group. This effect is more extreme as effect size and discrepancy in sample size are increased.

Comparing the random and weighted conditions, we can see that the random condition had a larger impact on classification accuracy of the DFA 50:50 condition than the weighted condition. However, overall the 50:50 condition appears to be impacted the least by misclassified data. The weighted condition had a stronger effect on classification accuracy of DFA than the random condition for the sample size discrepant conditions. The reduction in accuracy for the weighted condition ranged from 3-14 raw percentage points for the 25:75 condition and .9-15 raw percentage points for the 10:90 condition. In contrast, the reduction in accuracy for the random condition ranged from around 8 raw percentage points for the 25:75 condition and 1- 2 raw percentage points for the 10:90 condition. Sample size ratio also did not seem to make as strong an impact in the weighted condition, though this is likely due to the stronger effect of the misclassified data lowering the accuracy of DFA.

*Conclusion*

The results of these studies indicate that misclassification of training samples does have an impact on classification accuracy to a degree not previously understood or documented.  In addition, in line with previous results, increased sample size, effect size, and discrepancy in the ratio of sample sizes all lead to increases in overall classification accuracy.  We also learn that, consistent with Lachenbruch (1997) initial misclassification of groups (done at random) has relatively little impact on the classification accuracy of DFA.  In contrast, systematic misclassification of groups has a much larger impact on the classification accuracy of DFA. Neither type of misclassification affected the accuracy of FMM since initial classification is not part of the FMM.

In comparing accuracy between FMM and DFA, DFA displays higher classification accuracy in the majority of cases, whereas FMM displayed higher classification accuracy mainly in the 50:50 condition at the lower effect sizes.  The data also indicated the presence of an interaction between effect size and sample size ratio for FMM.  The increase in accuracy due to sample size ratio appears larger for the smaller sample sizes than for the larger sample sizes.

The introduction of weighted misclassified data had interesting effects on classification accuracy and direction of misclassification for DFA.  In general the same patterns regarding classification accuracy and interaction with other variables were observed, except weighted misclassified data produced more inaccuracy in classification than misclassification at random. It was also observed, that for the extreme cases of misclassification (especially for large effect sizes) in sample size discrepant conditions, weighted misclassified data causes the direction of misclassification to reverse.  It is likely, that misclassified data does not actually cause DFA to reverse its direction of misclassification, but rather, the extreme amount of misclassified data causes so many classification errors that the percents do not accurately represent how the data are being classified by the model.

Arguably the most noteworthy results of the paper are the direction of misclassification for finite mixture modeling and discriminant function analysis. When group size is discrepant, FMM and DFA tend to be biased toward making errors in different directions. For DFA, more classification errors are made classifying cases from the smaller group as being from the larger group. For FMM, more classification errors are made classifying cases from the larger group as being from the smaller group. Thus, even though the DFA achieves overall higher levels of classification accuracy in the majority of conditions, when group sizes are discrepant, FMM better captures the smaller group: a higher percentage of the cases from the smaller group are correctly classified. These results (found in study 1) were largely replicated in study 2 with weighted misclassified data (with the exception of DFA reversing direction at the extreme cases of misclassification). As will be discussed in the next section, these findings have important practical implications for determining when use of each technique is appropriate. A few studies have already documented the direction of misclassification for discriminant function analysis (Breckenridge, 2000; Lei & Koehly, 2003). However, the comparison with finite mixture modeling and implications for practical use are new contributions to the field.

*Discussion*

This paper shows that the relationship between initial misclassification of groups and classification accuracy differs depending on misclassification type, data and distributional characteristics, and analysis used. For finite mixture modeling the relationship is clear: initial misclassification of groups has no effect on classification accuracy, as finite mixture modeling does not use in any way initial knowledge of group status. For discriminant function analysis, where initial knowledge of group status is required, it is clear that there is a small effect when data are symmetrically misclassified at random and a larger effect when the data are symmetrically misclassified in a weighted fashion. We also learn that finite mixture modeling and discriminant function analysis tend to misclassify cases in opposite directions: finite mixture modeling misclassifying in favor of the smaller group, and discriminant function analysis misclassifying in favor of the larger group.

The results presented in this paper have practical implications for the decision of when to use each technique. Because of its demonstrated high levels of accuracy, discriminant function analysis may be the method of choice when the researcher is most interested in recovering the highest percentage correct. However, as the results of both study 1 and study 2 indicate, there are times when finite mixture modeling may provide a better alternative. In situations where the group sizes are approximately equal and expected effect size is low, finite mixture modeling has demonstrated higher levels of classification accuracy. In situations where group sizes are unequal, although it demonstrates lower classification accuracy overall, finite mixture modeling captures higher percents of the smaller group when sample size ratio is discrepant. It is easy to think of situations where identifying as many cases from a smaller target group is the most important goal of the analysis. For example, learning disabled students typically make up approximately 3-10% of the overall student population (Hallahan et al., 2007), and thus (when samples are representative of the population) we would see a sample size ratio of likely at most, 10:90. When screening students for learning problems, it is arguably more important to find as many of the at risk students as possible. Thus, an argument can be made that it is far more desirable to error in the direction of initially misclassifying more students as having learning difficulties who do not, than to overlook students who truly have a learning difficulties.  Follow up identification techniques can then be used to screen out students who do not have a disability (however, this is just one of many potential options for learning disability screening procedures). This is a situation where finite mixture modeling might provide a more desirable approach. Although the mixture model is likely to misclassify more non disabled individuals as learning disabled than the discriminant function analysis, individuals misclassified are going to have scores similar to those who are correctly classified.  Thus, these students are likely to be low achieving and would still benefit from the special interventions given to learning disabled students (Holden, 2007).  The same considerations could be made in the medical field.  When screening individuals for medical problems, it is arguably more important to diagnose an individual as having a medical problem (when they are healthy) than to miss the medical problem of an unhealthy individual.

Other considerations to make when selecting a statistical classification method might be the ratio of sample sizes and effect size expected.  If the researcher expects sample sizes to be approximately equal with a low effect size, a finite mixture model can provide a more accurate solution than a discriminant function analysis.

The robustness of discriminant function analysis to random misclassification in training data comes as welcome news. However, it is somewhat worrisome that weighted misclassified data poses such a threat to classification accuracy in discriminant function analysis, especially because it is the more realistic of the misclassification scenarios studies.  Discriminant function analysis is one of the most commonly used classification analysis procedures.  Although random data mistakes are easy to make, weighted misclassification due to implementation of cut score schemes is far more likely to occur in practice.  Even more worrisome, perhaps, is that researchers usually do not know the degree of misclassification which exists in our particular datasets. Thus, it is important to take possible effects of misclassification into account when interpreting results of a discriminant function analysis.

As a final note, there are several extensions to this line of research which could be beneficial. It is likely that the results of the present research would differ greatly if data were not restricted to symmetric misclassification. This paper provides strong evidence for the observation that when data are misclassified in a symmetric fashion there are small to moderate effects on the accuracy of discriminant function analysis. However, an extension may be to investigate how asymmetric misclassification impacts discriminant function analysis. When data are misclassified symmetrically, the group size ratio after misclassification stays more or less equal to the group size ratio of the true groups. What happens to classification accuracy when misclassification changes the group sizes? More research into this situation is warranted. It is also important to remember that the present studies only used one predictor variable in the analyses and restricted the situation to the two group case. Thus two other interesting extensions would be to multiple predictor variables, and multiple group cases.

Another interesting direction of research is looking into the consistency of classification for finite mixture modeling and discriminant function analysis. What can we learn if a case is initially classified as belonging to group 1 and is placed in group 1 by both finite mixture modeling and discriminant function analysis? What if it is initially classified as belonging to group 2 both discriminant function analysis and finite mixture modeling classify the case as belonging to group 1? It is possible that a combination of these approaches may be able to inform us about possible misclassification errors in our data. More research into this area is certainly warranted.

The results from the present studies provide strong implications for the practical use of DFA and FMM.  Research into misclassified data and misclassification analysis is very important for understanding when to use which classification method, and the implications of using one method over another.  A better understanding of these concepts will eventually lead to better accuracy in classification, and better accuracy in classification should be a goal for all areas of research that use classification methods.

References

Assakul, K., & Proctor, C. H.  (1967)  Testing independence in two-way contingency tables with data subject to misclassification.  *Psychometrika, 32*, 67-76.

Baker, F. B. (1979) Stability of two hierarchical grouping techniques.  Case 1:  Sensitivity to data errors.  *Journal of the American Statistical Association, 69,* 440-445.

Blashfield, R. K. (1976) Mixture model tests of cluster analysis:  Accuracy of four agglomerative hierarchical methods.  *Psychological Bulletin, 83,* 377-388.

Breckenridge, J. M.  (2000)  Validating cluster analysis:  Consistent replication and symmetry. *Multivariate Behavioral Research, 35,* 261-285.

Bross, I.  (1954)  Misclassification in 2x2 tables.  *Biometrics, 10,* 478-486.

Cohen, J. (1968) Weighted kappa:  Nominal scale agreement with provision for scaled disagreement or partial credit.  *Psychological Bulletin, 70,* 213-220.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences Second Edition.* Hillsdale, NJ: Lawrence Earlbaum Associates

deCraen, S., Commandeur, J. F., Frank, L. E., & Heiser, W. J.  (2006)  Effects of group size and lack of sphericity on the recovery of clusters in k-means cluster analysis. *Multivariate Behavioral Research, 41,* 127-145.

Dwyer, C. A. (1996)  Cut scores and testing: Statistics, judgment, truth and error.  *Psychological Assessment, 8,* 360-362.

Edelbrock, C.  (1979)  Mixture model tests of hierarchical clustering algorithms:  The problem of classifying everybody.  *Multivariate Behavioral Research, 14,* 367-384.

Finch, W. H., & Schneider, M. K.  (2006)  Misclassification rates for four methods of group classification:  Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. *Educational and Psychological Measurement, 66,* 240-257.

Flurry, B. (1997) *A First Course in Multivariate Statistics*. New York, New York: Springer-Verlag

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611:631.

Hallahan, D. P., Keller, C. E., Martinez, E. A., Gelman, J. A., & Fan, X. (2007) How variable are interstate prevalence rates of learning disabilities and other special education categories? A longitudinal comparison. *Council for Exceptional Children, 73,* 136-146.

Harrell, F. E., & Lee, K. I. (1985) A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.) *Biostatistics: Statistics in biomedical, public health and environmental sciences* (pp. 333-343), Amsterdam: Elsevier Science Publishers B. V.

Hastie, T., Tibshirani, R., & Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, New York: Springer-Verlag.

Holden, J. E. (2007) Identification of Learning Disabilities: Integrating approaches through use of statistical classification analysis. *Submitted Manuscript*

Katz, B. M., & McSweeney, M. (1979) Misclassification errors and categorical data analysis. *Journal of Experimental Education, 47,* 331-338.

Keogh, B. K. (2005) Revisiting classification and identification. *Learning Disability Quarterly, 28,* 100-102.

Klecka, W. R. (1980) *Discriminant Analysis*. Newbury Park, California: Sage Publications.

Kuiper, F. K., & Fisher, L. (1975) A Monte Carlo comparison of six clustering procedures. *Biometrics, 31,* 777-783.

Lachenbruch, P. A. (1966) Discriminant analysis when the initial samples are misclassified. *Technometrics*, *8,* 657-662.

Lathrop, R. L. (1986) Practical strategies for dealing with unreliability in competency Assessments. *Journal of Education Research, 70,* 234-237.

Lei, Pui-Wa, & Koehly, L. M. (2003) Linear discriminant analysis versus logistic regression: a comparison of classification errors in the two-group case. *The Journal of Experimental Education, 72,* 25-49.

Lubke, G., & Muthen, B. O. (2007) Performance of factor mixture models as a function of Model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14,* 26-47**.**

McLachlan, G. J., & Basford, K. E. (1998) *Mixture Models*. New York, New York: Marcel Dekker, Inc.

McLachlan, G., & Peel, D. (2000) *Finite Mixture Models*. New York, New York: John Wiley & Sons, Inc.

Meshbane, A., & Morris, J. D. (1996) Predictive discriminant analysis versus logistic regression in two-group classification problems. *Paper presented at the annual meeting of the American Educational Research Association, April 8, 1996, New York, NY.*

Milligan, Soon & Sokol (1983). The effect of cluster size, dimensionality, and the number of Clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis And Machine Intelligence, PAMI-5 (1),* 40-47.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/.

Rudner, L. M. (2001) Computing the expected proportions of misclassified examinees. *Practical Assessment, Research and Evaluation, 7, 1-7*

Rudner, L. M. (2001) Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7(14).*

Rudner, L. M. (2005) Expected classification accuracy. *Practical Assessment, Research and Evaluation, 10,* 1-4.

Tabachnick, B. G., & Fidell, L. S. (2007) *Using Multivariate Statistics Fifth Edition.* Boston, Massachusetts, Pearson Education Inc.

Van Ness, J. W., & Yang, J. J. (1998) Robust discriminant analysis: Training data breakdown point. *Journal of Statistical Planning and Inference, 67,* 67-83.

Venables, W. N., & Ripley, B. D. (2002) *Modern Applied Statistics With S Forth Edition.* New York, New York: Springer Science + Business Media LLC.

Zigler, E., & Phillips, L. (1961) Psychiatric diagnosis: A critique. *Journal of Abnormal and Social Psychology, 63,* 607-618.

Footnotes:

1. The term “group” is most appropriate when referring to a variable which is directly observable.  The term “class” is most appropriate when referring to a latent, or unobservable variable.
2. In Discriminant Function analysis, The covariance vectors ($\tilde{\mathbf{S}}$) need not be equivalent across groups.  More than two groups may also be used, though the focus of this paper restricts the number of groups to two.

Figure 1a.  Visual Representation of Random Misclassification

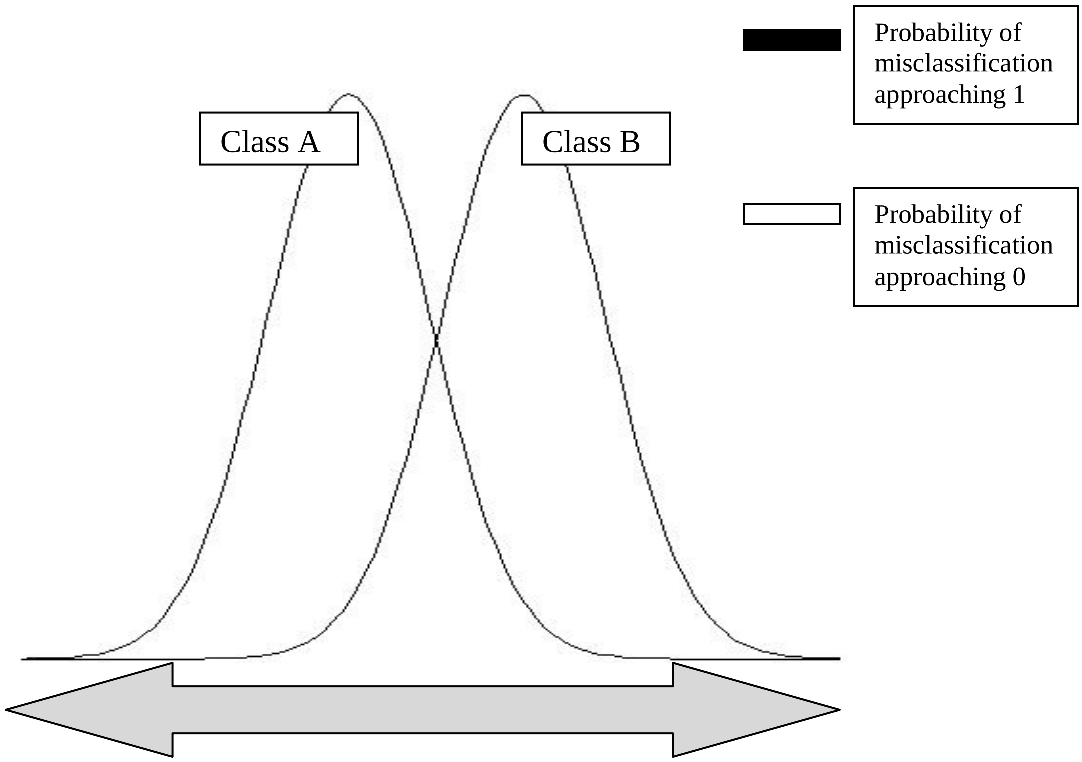


Figure 1b.  Visual Representation of Weighted Misclassification

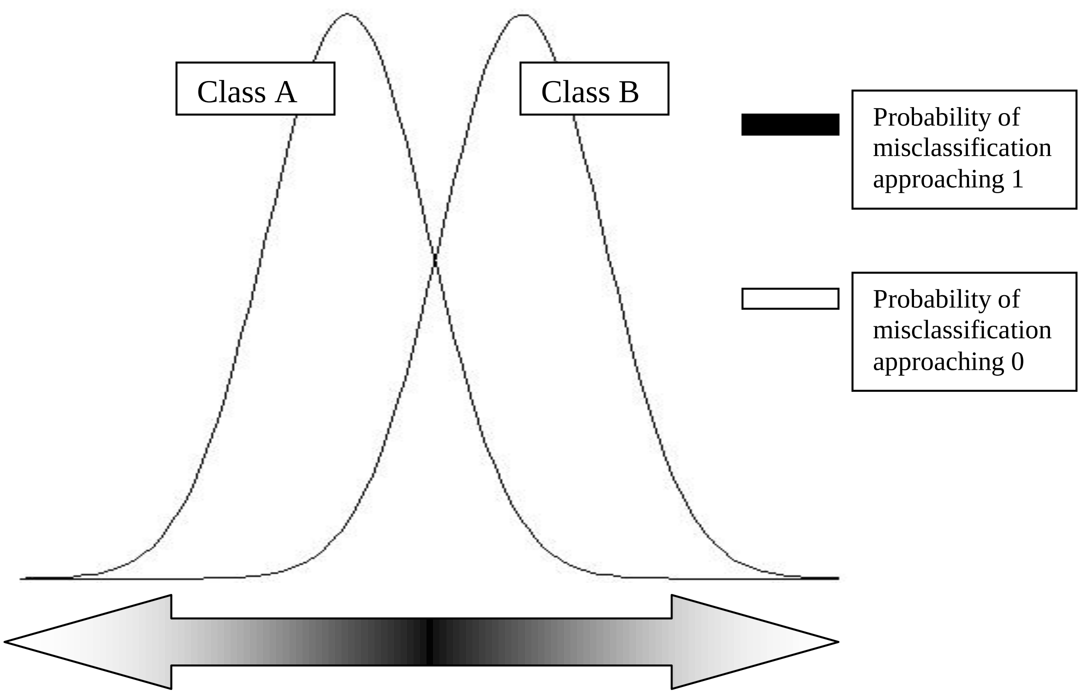

Table 1. Data Population Parameters for Study 1

| *Manipulated Variables* | |
|---|---|
| Percent Misclassified | 0%, 10%, 20%, 30% |
| Sample Size | 100, 500, 1000 |
| Sample Size Ratio | 50:50, 25:75, 10:90 |
| Standardized Mean Difference* | 0, .1, .2, .5, .8, 1.6 |

*For each condition, the population mean of group A was always 0. The population variance was held constant at 1 for both groups in all conditions.

Table 2. Percent Correctly Classified for Mixture Model and Discriminant Function with Sample Size 100.

| N=100 | | 100% | | 90% | | 80% | | 70% | |
|---|---|---|---|---|---|---|---|---|---|
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.538 | 0.500 | 0.538 | 0.499 | 0.538 | 0.500 | 0.538 | 0.500 |
| | 25:75 | 0.607 | 0.749 | 0.606 | 0.748 | 0.606 | 0.742 | 0.607 | 0.719 |
| | 10:90 | 0.662 | 0.900 | 0.662 | 0.900 | 0.662 | 0.898 | 0.662 | 0.888 |
| δ = .1 | 50:50 | 0.541 | 0.506 | 0.541 | 0.506 | 0.541 | 0.504 | 0.541 | 0.502 |
| | 25:75 | 0.608 | 0.749 | 0.608 | 0.748 | 0.608 | 0.742 | 0.608 | 0.720 |
| | 10:90 | 0.662 | 0.900 | 0.662 | 0.900 | 0.662 | 0.898 | 0.661 | 0.888 |
| δ = .2 | 50:50 | 0.548 | 0.527 | 0.547 | 0.518 | 0.547 | 0.514 | 0.547 | 0.509 |
| | 25:75 | 0.611 | 0.749 | 0.612 | 0.747 | 0.611 | 0.742 | 0.610 | 0.720 |
| | 10:90 | 0.664 | 0.900 | 0.664 | 0.900 | 0.663 | 0.898 | 0.662 | 0.888 |
| δ = .5 | 50:50 | 0.583 | 0.597 | 0.584 | 0.588 | 0.583 | 0.571 | 0.584 | 0.550 |
| | 25:75 | 0.626 | 0.748 | 0.625 | 0.747 | 0.626 | 0.743 | 0.624 | 0.725 |
| | 10:90 | 0.667 | 0.899 | 0.667 | 0.899 | 0.668 | 0.898 | 0.667 | 0.888 |
| δ = .8 | 50:50 | 0.626 | 0.654 | 0.627 | 0.650 | 0.627 | 0.637 | 0.627 | 0.609 |
| | 25:75 | 0.652 | 0.759 | 0.652 | 0.756 | 0.653 | 0.752 | 0.652 | 0.739 |
| | 10:90 | 0.680 | 0.899 | 0.677 | 0.899 | 0.680 | 0.898 | 0.680 | 0.889 |
| δ = 1.6 | 50:50 | 0.750 | 0.787 | 0.751 | 0.785 | 0.752 | 0.779 | 0.751 | 0.755 |
| | 25:75 | 0.771 | 0.833 | 0.772 | 0.829 | 0.771 | 0.820 | 0.772 | 0.802 |
| | 10:90 | 0.764 | 0.913 | 0.763 | 0.907 | 0.762 | 0.904 | 0.765 | 0.897 |

Note. N = sample size, δ = the standardized mean difference

Table 3. Percent Correctly Classified for Mixture Model and Discriminant Function with Sample Size 1000.

| N=1000 | | 100% | | 90% | | 80% | | 70% | |
|---|---|---|---|---|---|---|---|---|---|
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.513 | 0.500 | 0.513 | 0.500 | 0.513 | 0.500 | 0.513 | 0.500 |
| | 25:75 | 0.542 | 0.750 | 0.542 | 0.750 | 0.542 | 0.750 | 0.542 | 0.750 |
| | 10:90 | 0.567 | 0.900 | 0.566 | 0.900 | 0.566 | 0.900 | 0.565 | 0.900 |
| δ = .1 | 50:50 | 0.521 | 0.518 | 0.521 | 0.514 | 0.521 | 0.511 | 0.521 | 0.507 |
| | 25:75 | 0.544 | 0.750 | 0.545 | 0.750 | 0.545 | 0.750 | 0.544 | 0.750 |
| | 10:90 | 0.566 | 0.900 | 0.566 | 0.900 | 0.567 | 0.900 | 0.566 | 0.900 |
| δ = .2 | 50:50 | 0.538 | 0.540 | 0.539 | 0.538 | 0.539 | 0.533 | 0.539 | 0.525 |
| | 25:75 | 0.550 | 0.750 | 0.550 | 0.750 | 0.550 | 0.750 | 0.550 | 0.749 |
| | 10:90 | 0.567 | 0.900 | 0.567 | 0.900 | 0.567 | 0.900 | 0.567 | 0.900 |
| δ = .5 | 50:50 | 0.595 | 0.599 | 0.595 | 0.598 | 0.595 | 0.597 | 0.595 | 0.593 |
| | 25:75 | 0.584 | 0.751 | 0.584 | 0.751 | 0.584 | 0.750 | 0.584 | 0.749 |
| | 10:90 | 0.577 | 0.900 | 0.576 | 0.900 | 0.576 | 0.900 | 0.576 | 0.900 |
| δ= .8 | 50:50 | 0.651 | 0.655 | 0.651 | 0.655 | 0.651 | 0.654 | 0.650 | 0.652 |
| | 25:75 | 0.639 | 0.762 | 0.639 | 0.762 | 0.639 | 0.760 | 0.639 | 0.758 |
| | 10:90 | 0.603 | 0.900 | 0.602 | 0.900 | 0.602 | 0.900 | 0.603 | 0.900 |
| δ = 1.6 | 50:50 | 0.785 | 0.788 | 0.785 | 0.788 | 0.785 | 0.787 | 0.785 | 0.786 |
| | 25:75 | 0.818 | 0.835 | 0.819 | 0.834 | 0.819 | 0.833 | 0.819 | 0.829 |
| | 10:90 | 0.833 | 0.914 | 0.834 | 0.911 | 0.836 | 0.907 | 0.835 | 0.905 |

Note. N = sample size, δ = the standardized mean difference

Table 4. Percent Misclassified By Mixture Model and Discriminant Function under Random Condition

| N=100 | | 100% | | | | 90% | | | | 80% | | | | 70% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | |
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.229 | 0.250 | 0.233 | 0.250 | 0.232 | 0.250 | 0.230 | 0.250 | 0.230 | 0.248 | 0.232 | 0.252 | 0.232 | 0.251 | 0.230 | 0.250 |
| | 25:75 | 0.167 | 0.250 | 0.225 | 0.001 | 0.167 | 0.249 | 0.227 | 0.004 | 0.168 | 0.246 | 0.226 | 0.012 | 0.167 | 0.234 | 0.226 | 0.046 |
| | 10:90 | 0.068 | 0.100 | 0.270 | 0.000 | 0.068 | 0.100 | 0.270 | 0.000 | 0.068 | 0.100 | 0.270 | 0.002 | 0.068 | 0.099 | 0.270 | 0.014 |
| δ = .1 | 50:50 | 0.229 | 0.247 | 0.230 | 0.247 | 0.229 | 0.250 | 0.231 | 0.244 | 0.229 | 0.250 | 0.231 | 0.246 | 0.232 | 0.249 | 0.227 | 0.249 |
| | 25:75 | 0.166 | 0.249 | 0.226 | 0.001 | 0.166 | 0.249 | 0.226 | 0.004 | 0.167 | 0.245 | 0.226 | 0.013 | 0.166 | 0.234 | 0.225 | 0.046 |
| | 10:90 | 0.068 | 0.100 | 0.270 | 0.000 | 0.068 | 0.100 | 0.270 | 0.000 | 0.068 | 0.100 | 0.270 | 0.002 | 0.067 | 0.098 | 0.272 | 0.014 |
| δ = .2 | 50:50 | 0.227 | 0.237 | 0.226 | 0.236 | 0.229 | 0.241 | 0.224 | 0.241 | 0.226 | 0.242 | 0.227 | 0.244 | 0.227 | 0.246 | 0.225 | 0.245 |
| | 25:75 | 0.164 | 0.249 | 0.225 | 0.003 | 0.164 | 0.247 | 0.225 | 0.006 | 0.163 | 0.244 | 0.226 | 0.014 | 0.162 | 0.232 | 0.227 | 0.048 |
| | 10:90 | 0.067 | 0.100 | 0.269 | 0.000 | 0.067 | 0.100 | 0.268 | 0.001 | 0.067 | 0.100 | 0.270 | 0.002 | 0.067 | 0.098 | 0.271 | 0.014 |
| δ = .5 | 50:50 | 0.209 | 0.202 | 0.208 | 0.202 | 0.211 | 0.207 | 0.206 | 0.205 | 0.210 | 0.215 | 0.207 | 0.214 | 0.207 | 0.226 | 0.209 | 0.224 |
| | 25:75 | 0.147 | 0.238 | 0.227 | 0.013 | 0.149 | 0.237 | 0.226 | 0.016 | 0.149 | 0.233 | 0.225 | 0.024 | 0.148 | 0.221 | 0.228 | 0.054 |
| | 10:90 | 0.063 | 0.100 | 0.270 | 0.001 | 0.064 | 0.100 | 0.269 | 0.001 | 0.064 | 0.099 | 0.268 | 0.003 | 0.064 | 0.097 | 0.269 | 0.015 |
| δ = .8 | 50:50 | 0.188 | 0.173 | 0.186 | 0.173 | 0.186 | 0.175 | 0.187 | 0.175 | 0.187 | 0.180 | 0.186 | 0.183 | 0.187 | 0.196 | 0.186 | 0.195 |
| | 25:75 | 0.128 | 0.210 | 0.220 | 0.031 | 0.127 | 0.213 | 0.221 | 0.031 | 0.127 | 0.211 | 0.220 | 0.037 | 0.127 | 0.203 | 0.221 | 0.058 |
| | 10:90 | 0.058 | 0.097 | 0.262 | 0.003 | 0.057 | 0.098 | 0.266 | 0.003 | 0.058 | 0.098 | 0.262 | 0.005 | 0.058 | 0.095 | 0.262 | 0.016 |
| δ = 1.6 | 50:50 | 0.124 | 0.107 | 0.125 | 0.106 | 0.124 | 0.108 | 0.125 | 0.108 | 0.123 | 0.110 | 0.125 | 0.111 | 0.124 | 0.123 | 0.124 | 0.122 |
| | 25:75 | 0.086 | 0.116 | 0.143 | 0.052 | 0.087 | 0.122 | 0.142 | 0.049 | 0.087 | 0.129 | 0.142 | 0.051 | 0.087 | 0.137 | 0.140 | 0.061 |
| | 10:90 | 0.039 | 0.073 | 0.197 | 0.014 | 0.039 | 0.084 | 0.198 | 0.009 | 0.039 | 0.086 | 0.199 | 0.010 | 0.040 | 0.083 | 0.195 | 0.019 |

Note. When sample sizes are unequal, A is the larger group and B is the smaller group. Thus, A as B can be read as the percent of cases truly from the smaller group misclassified as belonging to the larger group.

Table 5. Percent Misclassified By Mixture Model and Discriminant Function under Random Condition

| N=1000 | | 100% | | | | 90% | | | | 80% | | | | 70% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | |
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.243 | 0.250 | 0.244 | 0.250 | 0.244 | 0.251 | 0.243 | 0.249 | 0.245 | 0.247 | 0.242 | 0.253 | 0.245 | 0.249 | 0.243 | 0.251 |
| | 25:75 | 0.143 | 0.250 | 0.315 | 0.000 | 0.143 | 0.250 | 0.315 | 0.000 | 0.143 | 0.250 | 0.315 | 0.000 | 0.143 | 0.250 | 0.314 | 0.000 |
| | 10:90 | 0.058 | 0.100 | 0.376 | 0.000 | 0.058 | 0.100 | 0.377 | 0.000 | 0.058 | 0.100 | 0.376 | 0.000 | 0.058 | 0.100 | 0.377 | 0.000 |
| δ = .1 | 50:50 | 0.240 | 0.241 | 0.239 | 0.241 | 0.239 | 0.242 | 0.240 | 0.244 | 0.240 | 0.245 | 0.239 | 0.244 | 0.239 | 0.244 | 0.240 | 0.249 |
| | 25:75 | 0.141 | 0.250 | 0.315 | 0.000 | 0.141 | 0.250 | 0.315 | 0.000 | 0.141 | 0.250 | 0.315 | 0.000 | 0.141 | 0.250 | 0.315 | 0.001 |
| | 10:90 | 0.057 | 0.100 | 0.376 | 0.000 | 0.057 | 0.100 | 0.377 | 0.000 | 0.057 | 0.100 | 0.376 | 0.000 | 0.057 | 0.100 | 0.377 | 0.000 |
| δ = .2 | 50:50 | 0.231 | 0.230 | 0.230 | 0.230 | 0.231 | 0.231 | 0.230 | 0.231 | 0.230 | 0.233 | 0.232 | 0.234 | 0.231 | 0.237 | 0.230 | 0.238 |
| | 25:75 | 0.135 | 0.250 | 0.315 | 0.000 | 0.134 | 0.250 | 0.316 | 0.000 | 0.134 | 0.250 | 0.317 | 0.000 | 0.134 | 0.249 | 0.317 | 0.001 |
| | 10:90 | 0.056 | 0.100 | 0.377 | 0.000 | 0.056 | 0.100 | 0.377 | 0.000 | 0.056 | 0.100 | 0.377 | 0.000 | 0.056 | 0.100 | 0.376 | 0.000 |
| δ = .5 | 50:50 | 0.203 | 0.201 | 0.202 | 0.201 | 0.203 | 0.201 | 0.202 | 0.201 | 0.203 | 0.201 | 0.202 | 0.202 | 0.202 | 0.202 | 0.203 | 0.205 |
| | 25:75 | 0.103 | 0.243 | 0.313 | 0.006 | 0.104 | 0.244 | 0.313 | 0.005 | 0.104 | 0.244 | 0.312 | 0.005 | 0.104 | 0.242 | 0.312 | 0.008 |
| | 10:90 | 0.050 | 0.100 | 0.374 | 0.000 | 0.312 | 0.005 | 0.104 | 0.244 | 0.050 | 0.100 | 0.375 | 0.000 | 0.049 | 0.100 | 0.375 | 0.000 |
| δ = .8 | 50:50 | 0.174 | 0.172 | 0.176 | 0.172 | 0.176 | 0.173 | 0.174 | 0.172 | 0.174 | 0.173 | 0.175 | 0.173 | 0.174 | 0.174 | 0.176 | 0.174 |
| | 25:75 | 0.085 | 0.209 | 0.276 | 0.029 | 0.084 | 0.215 | 0.277 | 0.024 | 0.084 | 0.218 | 0.276 | 0.021 | 0.085 | 0.219 | 0.277 | 0.022 |
| | 10:90 | 0.038 | 0.099 | 0.359 | 0.001 | 0.038 | 0.100 | 0.359 | 0.000 | 0.038 | 0.100 | 0.360 | 0.000 | 0.039 | 0.100 | 0.359 | 0.000 |
| δ =1.6 | 50:50 | 0.108 | 0.106 | 0.107 | 0.106 | 0.108 | 0.106 | 0.108 | 0.106 | 0.108 | 0.106 | 0.107 | 0.106 | 0.108 | 0.107 | 0.108 | 0.107 |
| | 25:75 | 0.088 | 0.114 | 0.093 | 0.051 | 0.088 | 0.117 | 0.093 | 0.048 | 0.088 | 0.122 | 0.093 | 0.045 | 0.088 | 0.128 | 0.093 | 0.043 |
| | 10:90 | 0.037 | 0.072 | 0.130 | 0.013 | 0.037 | 0.085 | 0.129 | 0.005 | 0.038 | 0.091 | 0.127 | 0.002 | 0.037 | 0.093 | 0.128 | 0.002 |

Note. When sample sizes are unequal, A is the larger group and B is the smaller group. Thus, A as B can be read as the percent of cases truly from the smaller group misclassified as belonging to the larger group.

Table 6. Matrix of Classification Decisions

| | | Classification Decision | |
|---|---|---|---|
| | | Target Group | Not Target Group |
| True in Population | Target Group | Correct | Miss |
| | Not Target Group | False Positive | Correct |

Table 7: Percent Correctly Classified for Mixture Model and Discriminant Function with Sample Size 100 under the weighted condition.

| N = 100 | | κ = 1 | | κ = .9 | | κ = .8 | | κ = .7 | |
|---|---|---|---|---|---|---|---|---|---|
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.534 | 0.502 | 0.534 | 0.499 | 0.534 | 0.500 | 0.534 | 0.499 |
| | 25:75 | 0.604 | 0.749 | 0.603 | 0.741 | 0.604 | 0.692 | 0.604 | 0.610 |
| | 10:90 | 0.658 | 0.900 | 0.660 | 0.896 | 0.659 | 0.859 | 0.660 | 0.751 |
| δ = .1 | 50:50 | 0.537 | 0.509 | 0.537 | 0.516 | 0.537 | 0.519 | 0.536 | 0.519 |
| | 25:75 | 0.605 | 0.749 | 0.604 | 0.738 | 0.604 | 0.691 | 0.606 | 0.620 |
| | 10:90 | 0.659 | 0.900 | 0.660 | 0.895 | 0.659 | 0.857 | 0.660 | 0.751 |
| δ = .2 | 50:50 | 0.544 | 0.526 | 0.544 | 0.534 | 0.543 | 0.537 | 0.543 | 0.538 |
| | 25:75 | 0.608 | 0.749 | 0.608 | 0.736 | 0.608 | 0.692 | 0.607 | 0.629 |
| | 10:90 | 0.659 | 0.900 | 0.659 | 0.895 | 0.662 | 0.856 | 0.660 | 0.753 |
| δ = .5 | 50:50 | 0.580 | 0.597 | 0.582 | 0.596 | 0.581 | 0.595 | 0.581 | 0.596 |
| | 25:75 | 0.625 | 0.749 | 0.623 | 0.737 | 0.623 | 0.705 | 0.623 | 0.661 |
| | 10:90 | 0.664 | 0.899 | 0.664 | 0.892 | 0.664 | 0.853 | 0.665 | 0.762 |
| δ = .8 | 50:50 | 0.627 | 0.655 | 0.625 | 0.652 | 0.626 | 0.651 | 0.626 | 0.652 |
| | 25:75 | 0.652 | 0.759 | 0.653 | 0.752 | 0.650 | 0.731 | 0.652 | 0.698 |
| | 10:90 | 0.676 | 0.899 | 0.678 | 0.891 | 0.678 | 0.856 | 0.677 | 0.776 |
| δ = 1.6 | 50:50 | 0.753 | 0.788 | 0.751 | 0.786 | 0.751 | 0.784 | 0.751 | 0.783 |
| | 25:75 | 0.771 | 0.833 | 0.770 | 0.829 | 0.771 | 0.819 | 0.770 | 0.796 |
| | 10:90 | 0.761 | 0.913 | 0.760 | 0.906 | 0.761 | 0.883 | 0.763 | 0.822 |

Note. N = sample size, δ = standardized mean difference

Table 8: Percent Correctly Classified for Mixture Model and Discriminant Function with Sample Size 1000 under the weighted condition.

| N = 1000 | | 100% | | 90% | | 80% | | 70% | |
|---|---|---|---|---|---|---|---|---|---|
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.512 | 0.500 | 0.512 | 0.500 | 0.512 | 0.500 | 0.512 | 0.500 |
| | 25:75 | 0.542 | 0.750 | 0.542 | 0.749 | 0.542 | 0.697 | 0.542 | 0.609 |
| | 10:90 | 0.566 | 0.900 | 0.565 | 0.900 | 0.566 | 0.870 | 0.565 | 0.750 |
| δ = .1 | 50:50 | 0.521 | 0.517 | 0.521 | 0.520 | 0.521 | 0.520 | 0.521 | 0.520 |
| | 25:75 | 0.544 | 0.750 | 0.544 | 0.746 | 0.543 | 0.694 | 0.545 | 0.618 |
| | 10:90 | 0.566 | 0.900 | 0.565 | 0.900 | 0.566 | 0.867 | 0.567 | 0.752 |
| δ = .2 | 50:50 | 0.538 | 0.540 | 0.539 | 0.540 | 0.538 | 0.540 | 0.538 | 0.540 |
| | 25:75 | 0.550 | 0.750 | 0.550 | 0.743 | 0.549 | 0.695 | 0.549 | 0.628 |
| | 10:90 | 0.566 | 0.900 | 0.568 | 0.899 | 0.567 | 0.864 | 0.567 | 0.753 |
| δ = .5 | 50:50 | 0.595 | 0.599 | 0.595 | 0.599 | 0.595 | 0.599 | 0.595 | 0.598 |
| | 25:75 | 0.584 | 0.751 | 0.584 | 0.741 | 0.583 | 0.708 | 0.583 | 0.662 |
| | 10:90 | 0.575 | 0.900 | 0.577 | 0.897 | 0.576 | 0.859 | 0.576 | 0.763 |
| δ = .8 | 50:50 | 0.650 | 0.655 | 0.650 | 0.655 | 0.651 | 0.655 | 0.650 | 0.655 |
| | 25:75 | 0.639 | 0.762 | 0.640 | 0.756 | 0.639 | 0.735 | 0.639 | 0.700 |
| | 10:90 | 0.603 | 0.900 | 0.603 | 0.895 | 0.602 | 0.861 | 0.601 | 0.777 |
| δ = 1.6 | 50:50 | 0.785 | 0.788 | 0.785 | 0.788 | 0.785 | 0.788 | 0.785 | 0.788 |
| | 25:75 | 0.819 | 0.835 | 0.819 | 0.832 | 0.818 | 0.822 | 0.818 | 0.801 |
| | 10:90 | 0.832 | 0.915 | 0.833 | 0.911 | 0.834 | 0.888 | 0.834 | 0.825 |

Note. N = sample size, δ = standardized mean difference

Table 9. Percent Misclassified By Mixture Model and Discriminant Function with sample size 100 under the Weighted Condition

| N = 100 | | 100% | | | | 90% | | | | 80% | | | | 70% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | |
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.232 | 0.250 | 0.234 | 0.250 | 0.233 | 0.250 | 0.232 | 0.251 | 0.233 | 0.250 | 0.233 | 0.250 | 0.234 | 0.250 | 0.232 | 0.251 |
| | 25:75 | 0.166 | 0.250 | 0.231 | 0.001 | 0.166 | 0.243 | 0.229 | 0.021 | 0.166 | 0.213 | 0.231 | 0.109 | 0.166 | 0.173 | 0.230 | 0.232 |
| | 10:90 | 0.067 | 0.100 | 0.276 | 0.000 | 0.068 | 0.099 | 0.273 | 0.007 | 0.068 | 0.093 | 0.274 | 0.064 | 0.067 | 0.077 | 0.275 | 0.202 |
| δ = .1 | 50:50 | 0.231 | 0.246 | 0.232 | 0.245 | 0.231 | 0.242 | 0.232 | 0.243 | 0.232 | 0.242 | 0.231 | 0.240 | 0.233 | 0.241 | 0.230 | 0.241 |
| | 25:75 | 0.165 | 0.249 | 0.230 | 0.002 | 0.166 | 0.239 | 0.230 | 0.028 | 0.165 | 0.205 | 0.230 | 0.117 | 0.164 | 0.164 | 0.231 | 0.232 |
| | 10:90 | 0.067 | 0.100 | 0.274 | 0.000 | 0.067 | 0.099 | 0.275 | 0.009 | 0.068 | 0.091 | 0.272 | 0.067 | 0.068 | 0.074 | 0.271 | 0.204 |
| δ = .2 | 50:50 | 0.229 | 0.237 | 0.227 | 0.236 | 0.229 | 0.232 | 0.228 | 0.233 | 0.229 | 0.230 | 0.227 | 0.231 | 0.231 | 0.231 | 0.225 | 0.231 |
| | 25:75 | 0.163 | 0.248 | 0.231 | 0.003 | 0.162 | 0.233 | 0.232 | 0.036 | 0.162 | 0.196 | 0.229 | 0.124 | 0.163 | 0.156 | 0.229 | 0.230 |
| | 10:90 | 0.067 | 0.100 | 0.274 | 0.000 | 0.066 | 0.098 | 0.275 | 0.010 | 0.067 | 0.089 | 0.273 | 0.072 | 0.067 | 0.071 | 0.275 | 0.203 |
| δ = .5 | 50:50 | 0.211 | 0.202 | 0.207 | 0.201 | 0.209 | 0.202 | 0.210 | 0.202 | 0.209 | 0.203 | 0.210 | 0.202 | 0.207 | 0.202 | 0.210 | 0.202 |
| | 25:75 | 0.147 | 0.238 | 0.231 | 0.014 | 0.146 | 0.208 | 0.231 | 0.059 | 0.147 | 0.167 | 0.230 | 0.136 | 0.147 | 0.131 | 0.230 | 0.218 |
| | 10:90 | 0.063 | 0.100 | 0.273 | 0.001 | 0.063 | 0.095 | 0.271 | 0.017 | 0.063 | 0.081 | 0.273 | 0.079 | 0.062 | 0.061 | 0.274 | 0.202 |
| δ = .8 | 50:50 | 0.184 | 0.173 | 0.189 | 0.173 | 0.188 | 0.174 | 0.187 | 0.173 | 0.189 | 0.175 | 0.185 | 0.174 | 0.189 | 0.174 | 0.185 | 0.173 |
| | 25:75 | 0.127 | 0.209 | 0.224 | 0.032 | 0.125 | 0.174 | 0.225 | 0.076 | 0.125 | 0.139 | 0.224 | 0.134 | 0.127 | 0.109 | 0.222 | 0.202 |
| | 10:90 | 0.057 | 0.097 | 0.268 | 0.004 | 0.058 | 0.089 | 0.267 | 0.023 | 0.057 | 0.072 | 0.266 | 0.085 | 0.057 | 0.050 | 0.267 | 0.197 |
| δ = 1.6 | 50:50 | 0.125 | 0.107 | 0.125 | 0.107 | 0.125 | 0.107 | 0.126 | 0.107 | 0.123 | 0.108 | 0.127 | 0.108 | 0.124 | 0.109 | 0.126 | 0.108 |
| | 25:75 | 0.086 | 0.115 | 0.144 | 0.052 | 0.085 | 0.097 | 0.144 | 0.075 | 0.087 | 0.078 | 0.141 | 0.107 | 0.086 | 0.059 | 0.142 | 0.151 |
| | 10:90 | 0.039 | 0.073 | 0.200 | 0.015 | 0.039 | 0.060 | 0.200 | 0.035 | 0.039 | 0.042 | 0.202 | 0.084 | 0.038 | 0.025 | 0.200 | 0.173 |

Note. When sample sizes are unequal, A is the larger group and B is the smaller group. Thus, A as B can be read as the percent of cases truly from the smaller group misclassified as belonging to the larger group.

Table 10. Percent Misclassified By Mixture Model and Discriminant Function with sample size 1000 under the Weighted Condition

| N = 1000 | | 100% | | | | 90% | | | | 80% | | | | 70% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | | A as B | | B as A | |
| | A:B | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA | FMM | DFA |
| δ = 0 | 50:50 | 0.244 | 0.250 | 0.244 | 0.250 | 0.244 | 0.250 | 0.244 | 0.250 | 0.244 | 0.250 | 0.244 | 0.250 | 0.243 | 0.250 | 0.245 | 0.250 |
| | 25:75 | 0.143 | 0.250 | 0.315 | 0.000 | 0.143 | 0.249 | 0.314 | 0.002 | 0.143 | 0.224 | 0.315 | 0.079 | 0.143 | 0.180 | 0.314 | 0.211 |
| | 10:90 | 0.058 | 0.100 | 0.377 | 0.000 | 0.058 | 0.100 | 0.377 | 0.000 | 0.058 | 0.096 | 0.376 | 0.033 | 0.058 | 0.081 | 0.377 | 0.169 |
| δ = .1 | 50:50 | 0.241 | 0.241 | 0.238 | 0.241 | 0.240 | 0.240 | 0.239 | 0.240 | 0.240 | 0.240 | 0.239 | 0.240 | 0.239 | 0.240 | 0.240 | 0.240 |
| | 25:75 | 0.141 | 0.250 | 0.315 | 0.000 | 0.141 | 0.247 | 0.315 | 0.007 | 0.140 | 0.214 | 0.316 | 0.092 | 0.141 | 0.171 | 0.314 | 0.211 |
| | 10:90 | 0.057 | 0.100 | 0.377 | 0.000 | 0.057 | 0.100 | 0.378 | 0.001 | 0.057 | 0.095 | 0.376 | 0.038 | 0.057 | 0.078 | 0.376 | 0.170 |
| δ = .2 | 50:50 | 0.230 | 0.230 | 0.232 | 0.230 | 0.231 | 0.230 | 0.231 | 0.230 | 0.231 | 0.230 | 0.231 | 0.230 | 0.231 | 0.230 | 0.231 | 0.230 |
| | 25:75 | 0.134 | 0.250 | 0.316 | 0.000 | 0.134 | 0.243 | 0.316 | 0.014 | 0.134 | 0.204 | 0.318 | 0.101 | 0.134 | 0.163 | 0.317 | 0.209 |
| | 10:90 | 0.056 | 0.100 | 0.378 | 0.000 | 0.056 | 0.100 | 0.376 | 0.001 | 0.056 | 0.093 | 0.377 | 0.043 | 0.056 | 0.075 | 0.377 | 0.171 |
| δ = .5 | 50:50 | 0.202 | 0.201 | 0.203 | 0.201 | 0.202 | 0.201 | 0.202 | 0.201 | 0.202 | 0.201 | 0.202 | 0.201 | 0.202 | 0.201 | 0.203 | 0.201 |
| | 25:75 | 0.103 | 0.243 | 0.313 | 0.007 | 0.104 | 0.214 | 0.312 | 0.045 | 0.103 | 0.173 | 0.314 | 0.118 | 0.104 | 0.137 | 0.313 | 0.201 |
| | 10:90 | 0.050 | 0.100 | 0.375 | 0.000 | 0.049 | 0.098 | 0.374 | 0.005 | 0.049 | 0.085 | 0.375 | 0.055 | 0.049 | 0.065 | 0.375 | 0.172 |
| δ = .8 | 50:50 | 0.175 | 0.172 | 0.175 | 0.172 | 0.175 | 0.172 | 0.174 | 0.172 | 0.175 | 0.173 | 0.175 | 0.172 | 0.175 | 0.172 | 0.175 | 0.172 |
| | 25:75 | 0.084 | 0.209 | 0.277 | 0.029 | 0.085 | 0.177 | 0.276 | 0.067 | 0.084 | 0.144 | 0.277 | 0.122 | 0.084 | 0.113 | 0.277 | 0.187 |
| | 10:90 | 0.038 | 0.099 | 0.359 | 0.001 | 0.038 | 0.092 | 0.359 | 0.012 | 0.038 | 0.075 | 0.360 | 0.064 | 0.038 | 0.053 | 0.361 | 0.170 |
| δ = 1.6 | 50:50 | 0.107 | 0.106 | 0.108 | 0.106 | 0.108 | 0.106 | 0.107 | 0.106 | 0.108 | 0.106 | 0.108 | 0.106 | 0.107 | 0.106 | 0.108 | 0.106 |
| | 25:75 | 0.089 | 0.114 | 0.093 | 0.051 | 0.089 | 0.098 | 0.093 | 0.070 | 0.088 | 0.079 | 0.094 | 0.098 | 0.088 | 0.061 | 0.093 | 0.138 |
| | 10:90 | 0.036 | 0.072 | 0.132 | 0.013 | 0.037 | 0.061 | 0.130 | 0.027 | 0.037 | 0.044 | 0.130 | 0.068 | 0.037 | 0.027 | 0.129 | 0.148 |

Note. When sample sizes are unequal, A is the larger group and B is the smaller group. Thus, A as B can be read as the percent of cases truly from the smaller group misclassified as belonging to the larger group.