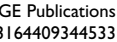
Educational and Psychological Measurement
70(1) 36–55

DOI: 10.1177/0013164409344533
http://epm.sagepub.com



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

Jocelyn E. Holden[1] and Ken Kelley[2]

Abstract

Classification procedures are common and useful in behavioral, educational, social, and managerial research. Supervised classification techniques such as discriminant function analysis assume training data are perfectly classified when estimating parameters or classifying. In contrast, unsupervised classification techniques such as finite mixture models (FMM) do not require, or even use if available, knowledge of group status to estimate parameters or classifying. This study investigates the impact of two types of misclassification errors on the classification accuracy of discriminant function analysis (both linear [LDA] and quadratic [QDA]) and FMM for two groups with a single predictor. Analytic and Monte Carlo results are provided for a variety of misclassification scenarios to investigate the performance of the two methods. Discriminant function techniques recovered the highest overall percentages of correctly classified data, whereas FMM captured higher percentages of the smaller group when group sizes are unequal. LDA marginally outperformed QDA under misclassified conditions.

Keywords

classification, misclassification, linear discriminant function analysis, quadratic discriminant function analysis, mixture model, training data

Classification of individuals into nonoverlapping groups is regularly used in the behavioral, educational, social, and managerial research and practice, as well as in

[1]Indiana University, Bloomington, USA
[2]University of Notre Dame, IN, USA

Corresponding Author:
Ken Kelley, Department of Management, University of Notre Dame, Notre Dame, IN 46556, USA
E-mail: kkelley@nd.edu

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 when 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 be observed (e.g., gender, grade level, employment status) or latent (e.g., learning disabled, depressed, or alcoholic). The term *group* is most appropriate when referring to a variable that is directly observable or otherwise known. The term *class* is most appropriate when referring to a latent or unobservable variable. When class is latent, or unobservable, misclassification errors are almost unavoidable because it is generally not possible to know an individual's true class with certainty. 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, or 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. Apart 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 used. Further, evidence from simulation studies reveals statistical classification methods have varying degrees of classification accuracy under a variety of different situations. For example, it has been found 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, Commandeur, Frank, & Heiser, 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 (DFA) 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 have shown 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.

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 DFA). 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 have misclassification errors? How much does initial misclassification of training data affect the ability of classification analysis schemes to accurately recover groups?

For chi-square analyses it has been found that misclassification of data categories does not affect the validity of the test of significance, although it may reduce the power of the test (Bross, 1954; Assakul & Proctor, 1967; Katz & McSweeney, 1979). For DFA, however, initial misclassification of training data does have an impact on classification accuracy (Lachenbruch, 1966, 1974, 1979; McLachlan, 1972; Chhikara & McKeon, 1984; Grayson, 1987). The distinction between random and nonrandom misclassification has been demonstrated to be an important distinction in terms of classification accuracy. Several studies have demonstrated that when discriminant function training data are misclassified at random, they have a larger impact on classification accuracy than misclassification occurring in a nonrandom fashion (Lachenbruch, 1966, 1974, 1979; McLachlan, 1972; Chhikara & McKeon, 1984). In particular, when misclassified training data are incorporated into a DFA, the percentage of correctly classified cases is systematically underestimated (Lachenbruch, 1966, 1974, 1979; McLachlan, 1972).

When comparing techniques under misclassified data conditions, it has been found that linear discriminant function analysis (LDA) is less affected than quadratic discriminant function analysis (QDA). In particular, as the rate of misclassified cases increases, the sample covariance matrices increasingly differ systematically from the population parameters (Lachenbruch, 1979).

These studies provide substantial evidence for the negative effects of misclassified training data on LDA and QDA. However, these studies were limited in the scope of variables investigated. For each of these studies, the ratio of group sizes was assumed equal, thus ignoring the question of whether the ratio of group sizes changes the impact of misclassification.

The purpose of the present research is to discern the effect initially misclassifying data has on the effectiveness of the two group case of LDA, QDA, and finite mixture modeling (FMM) under various data and distributional characteristics and to provide comparisons between the methods. The rationale for choosing these comparison methods is twofold: Comparing accuracy between the linear and quadratic forms of DFA under misclassified conditions provides a replication under different conditions of Lachenbruch's (1979) findings. Also, because QDA and the FMM used in this study both assume unequal variances, a fair comparison can be made between supervised classification and unsupervised classification techniques. Because LDA and QDA are based on more information than FMM, we make the following hypotheses:

Hypothesis 1: When training data is perfectly classified, the discriminant function models 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 models to provide accurate classifications will continue to decrease while the finite mixture model accuracy will remain unchanged.

Based on results from the existing literature, a third hypothesis can be made:

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 models these characteristics may interact with initially misclassified data proportions to produce poorer misclassification.

Conceptual Examination of Discriminant Function Analysis

An examination of the mathematics involved in LDA, QDA, and FMM also dictates when effects of misclassification should be observed. Recall that the procedure for DFA involves choosing the combination (linear or quadratic depending on whether equal variances are assumed) 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 that classifies each individual in the group to which their specific discriminant function score is most similar. Let $\mathbf{a}$ be a vector of coefficients, $\mathbf{S}$ be the unbiased pooled estimate of the population covariance matrix (which in the case of LDA assumes homogeneity of population covariance matrices: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the common population covariance matrix), $\overline{\mathbf{x}}_j$ be the mean vector of length p, where p is the number of variables for the cases in group j, and $D(\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2)$ be the multivariate distance between $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$, which is defined as

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

When $D(\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2)$ is maximized by a particular vector $\mathbf{a}$, the resulting $\mathbf{a}$ vector becomes the vector of discriminant function coefficients. Conceptually, the multivariate distance between $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ is the maximum of the univariate distances between $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ (Flury, 1997). 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

$$\overline{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n_i} x_{ij}, \tag{2}$$

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. When group status (j) is incorrect for certain individuals, the centroids from Equation 2 are less likely to represent the true population values. Thus, any classifications made from these biased centroids will tend to contain some degree of classification error. Based on the DFA equations, the accuracy of discriminant function solutions should decrease as the amount of misclassified data is increased, because misclassification will tend to bias the mean vectors.

Conceptual Examination of Finite Mixture Modeling

Unlike DFA, the definitional equations for FMM show that initial misclassification of group status will have no effect on the results. Finite mixture models are unsupervised classification models that can be used whether or not the true group status is initially known. In the present study, the FMM was used to recover the two groups known to exist in the data. However, FMM are highly customizable and can be implemented for any number of groups, or can be used to recover the number of groups which best fits the data. Thus FMM can be customized for use with either homogeneous or heterogeneous variances, different distributional shapes, and so on. The sample composite distribution function based on the finite mixture model has the form

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

where $\hat{\pi}_m$ are the sample mixing proportions ($\sum_{m=1}^{M} \hat{\pi}_m = 1$), and represents the normal distribution function with sample mean $\bar{\mathbf{x}}_m$ and covariance matrix $\mathbf{S}_m$, also called component distributions because they are the distributions that 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:

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

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

Thus, finite mixture models are based only on information in the data and by the number of groups specified. All other parameters are estimated by the model under the particular assumptions. The finite mixture model used for the present study assumes normal distributions with their own variance for each component density.

Note that in the case of the FMM, no initial group status is needed. The accuracy of initial classification, if any, is not an issue. Thus, 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 LDA, QDA, and FMM to classify accurately when training data is misclassified above and beyond any previously reported "main effects" that may exist. Such characteristics include sample size, group size ratio, distance between group means, and variance of distributions. Based on these results, our third hypothesis is also justified.

Because misclassification problems can take different forms, we investigate the effects of two different types of misclassification. The two types of misclassification examined were random and nonrandom. *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. *Nonrandom 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 are initially misclassified at random to varying degrees. Once generated, data were analyzed with LDA, QDA, and FMM in an attempt to recover the true groups and determine the effectiveness of each method.

Methods

Data generation. Generation and analysis of misclassified data was accomplished using the R statistical software program (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 480 ($3 \times 4 \times 2 \times 5 \times 4$) 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 (0.2), medium (0.5), and large (0.8) effect, as well as "very large," which we operationalized as 1.6, twice the size of large (see Kelley & Rausch, 2006, for a review of the standardized mean difference and some of its properties). To test for the effects due to having a larger or smaller mean, five different sample size ratios were tested: 50:50 (where both groups are of equal size); two ratios where the smaller group has the smaller mean (25:75 and 10:90), and two ratios where the smaller group has the larger mean (75:25 and 90:10).

Table 1. Data Population Parameters

Data conditions	
True groups	2
Population variance (within each group)	1
Manipulated variables	
Statistical analysis	LDA, QDA, FMM
Percentage misclassified	0%, 10%, 20%, 30%
Sample size	100, 1000
Sample size ratio	10:90, 25:75, 50:50, 75:25, 90:10
Standardized mean difference (δ)	0.20, 0.50, 0.80, 1.6

Note: 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. LDA = linear discriminant function analysis; QDA = quadratic discriminant function analysis; FMM = finite mixture model.

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. Data were also limited to one predictor variable to better discern the effects of each factor. 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 randomly, the following procedure was used: First, two classes of data with specified means and standard deviations were generated (see Table 1). In Study 1, misclassification was at random, making every individual just as likely to be misclassified as every other individual. For each individual's score, a random number between 0 and 1 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, for 90% of the data to be correctly classified, individuals with random numbers <0.10 would be misclassified thus ensuring approximately 10% of the cases would be misclassified on every iteration and the average misclassification across all iterations would be 10%. For more detailed information on data generation, see Appendix B.

Analyses. After data were misclassified, a FMM assuming unequal variances, LDA and QDA were performed on the misclassified data. For FMM separate variance parameters and prior probabilities were estimated for each group. At the completion of each analysis, the classification results were compared to known classes and a percentage of correctly classified cases was recorded. Also calculated are the smaller group misclassification rate and the larger group misclassification rate. Analysis of these percents allows investigation into whether, and under what conditions, the methods may be biased toward misclassification in one direction. A total of 10,000 replications of this procedure were executed for each of the 480 simulation conditions. R code is available from authors on request.

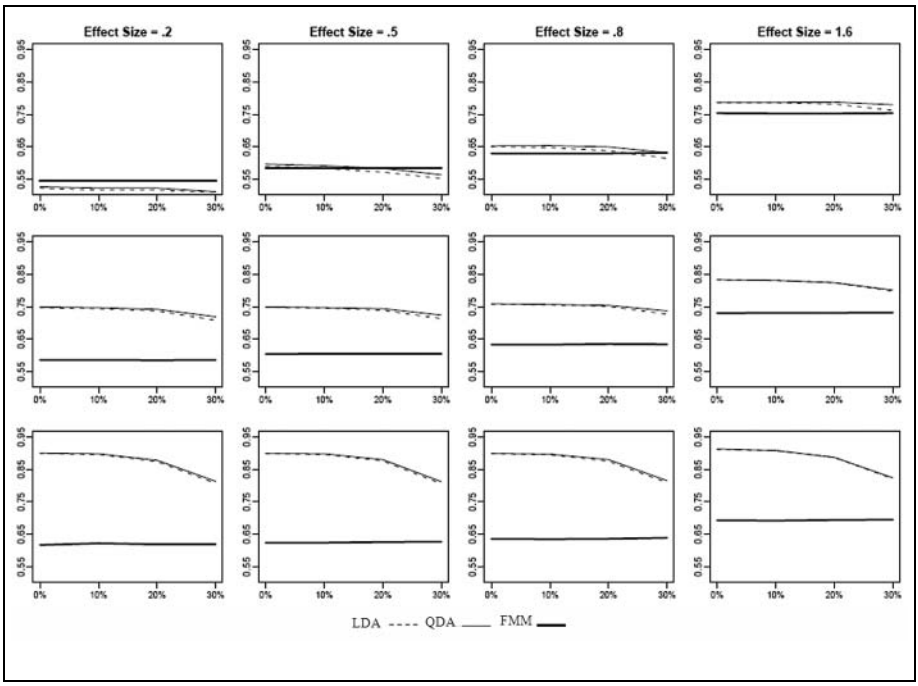


Figure 1. Percentage correct for Study 1 ($N = 100$)
Note: The *Y* axis presents the percentage of misclassified data. The *X* axis is the overall percentage correct achieved by the models. LDA = linear discriminant function analysis (dashed line); QDA = quadratic discriminant function analysis (thin solid line); FMM = finite mixture model (thick solid line).

Results

The percentages of correctly classified cases for each classification method are displayed in graph form in Figure 1. Although two sample sizes were tested, $N = 100$ and $N = 1{,}000$, for space considerations results are only displayed for the smallest sample. Relationships between variables were essentially the same for both sample size conditions. Readers can request full set of results from the authors. No biasing effect was found for groups having a larger or smaller mean: Percentage correct was exactly the same for the 10:90 condition as for the 90:10 condition. The same larger and smaller group classification rates were also found (Table 2). Thus, results for sample size ratio will only be displayed for 10:90, 25:75, and 50:50 to avoid redundant information.

Consistent with previous research it was found that as effect size and sample size increase, the ability of all three classification analyses to correctly classify data increases. LDA and QDA outperformed FMM in the majority of conditions overall. However, FMM showed marginally higher levels of classification accuracy in the 50:50 condition at the lower effect sizes. Consistent with Lachenbruch (1979),

Table 2. Small Group and Large Group Misclassification Rates for Random $N = 100$ Condition

			Percentage misclassified											
			0%			10%			20%			30%		
		A:B	FMM	LDA	QDA	FMM	LDA	QDA	FMM	LDA	QDA	FMM	LDA	QDA
Small group misclassification rate	$\delta = 0.2$	10:90	0.059	0.100	0.100	0.060	0.100	0.099	0.060	0.097	0.097	0.060	0.089	0.088
		25:75	0.146	0.249	0.248	0.146	0.247	0.246	0.145	0.244	0.241	0.145	0.232	0.227
		50:50	0.228	0.236	0.239	0.227	0.239	0.240	0.227	0.241	0.242	0.227	0.244	0.245
	$\delta = 0.5$	10:90	0.055	0.100	0.099	0.056	0.099	0.099	0.056	0.097	0.096	0.055	0.088	0.087
		25:75	0.129	0.238	0.238	0.129	0.238	0.236	0.129	0.235	0.232	0.129	0.224	0.218
		50:50	0.208	0.201	0.204	0.208	0.204	0.208	0.208	0.208	0.213	0.210	0.218	0.224
	$\delta = 0.8$	10:90	0.049	0.098	0.098	0.049	0.098	0.097	0.049	0.095	0.093	0.049	0.087	0.085
		25:75	0.111	0.210	0.212	0.109	0.214	0.213	0.109	0.213	0.209	0.110	0.206	0.201
		50:50	0.186	0.173	0.175	0.187	0.173	0.177	0.186	0.175	0.182	0.184	0.184	0.192
	$\delta = 1.6$	10:90	0.032	0.073	0.074	0.032	0.081	0.075	0.032	0.083	0.074	0.032	0.078	0.070
		25:75	0.065	0.115	0.114	0.064	0.122	0.122	0.064	0.130	0.128	0.064	0.139	0.134
		50:50	0.123	0.107	0.107	0.124	0.107	0.107	0.124	0.106	0.109	0.123	0.110	0.118
Large group misclassification rate	$\delta = 0.2$	10:90	0.323	0.000	0.001	0.318	0.003	0.005	0.321	0.025	0.030	0.321	0.098	0.105
		25:75	0.268	0.003	0.005	0.268	0.005	0.010	0.269	0.015	0.022	0.269	0.049	0.065
		50:50	0.225	0.236	0.238	0.227	0.239	0.242	0.227	0.241	0.243	0.227	0.244	0.246
	$\delta = 0.5$	10:90	0.320	0.001	0.002	0.320	0.003	0.006	0.319	0.023	0.028	0.318	0.100	0.107
		25:75	0.267	0.013	0.014	0.266	0.015	0.018	0.265	0.022	0.029	0.266	0.052	0.068
		50:50	0.208	0.201	0.205	0.209	0.204	0.209	0.208	0.208	0.215	0.207	0.218	0.223
	$\delta = 0.8$	10:90	0.315	0.004	0.004	0.316	0.005	0.008	0.315	0.025	0.032	0.312	0.099	0.107
		25:75	0.255	0.031	0.030	0.258	0.029	0.031	0.256	0.033	0.040	0.257	0.057	0.072
		50:50	0.185	0.173	0.174	0.184	0.173	0.175	0.184	0.175	0.180	0.184	0.184	0.193
	$\delta = 1.6$	10:90	0.275	0.014	0.015	0.276	0.011	0.017	0.274	0.030	0.040	0.273	0.097	0.109
		25:75	0.205	0.052	0.053	0.205	0.048	0.047	0.205	0.046	0.047	0.204	0.060	0.068
		50:50	0.124	0.107	0.107	0.122	0.106	0.107	0.123	0.106	0.109	0.124	0.110	0.119

LDA = linear discriminant function analysis; QDA = quadratic discriminant function analysis; FMM = finite mixture model.

LDA slightly outperformed QDA under misclassified data conditions. When group sizes were equal, the difference between LDA and QDA was larger, with the results becoming more similar as sample size ratio increased. The difference between LDA and QDA also increased as the percentage of misclassified data increased. Differences between LDA and QDA reduce as sample size increases. However, in comparison with FMM, the difference between LDA and QDA is only marginal. 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 LDA and QDA to classify correctly increases dramatically. For FMM, there is an increase in classification accuracy due to sample size ratio at the lower effect sizes. As effect size increases the classification between the sample size ratios becomes more similar, and then reverses direction. That is, at the higher effect sizes there is a decrease in classification accuracy as sample size ratio increases. Introduction of misclassified cases did not affect the overall finite mixture model classification in any way as initial classification is not part of the model.

As expected, no effect of misclassified data on the misclassification direction of FMM was observed. However, for misclassification for LDA and QDA there are marginal effects of misclassified data. As the percentage of misclassified data increases, the percentage of the smaller group misclassified as the larger group decreases, and the percentage of the larger group misclassified as the smaller group increases. Interestingly, although QDA generally has a lower overall classification rate, it misclassifies less of the small group than LDA. Although the difference is very small, it is consistent and an important finding. Besides the fact that QDA misclassifies less of the small group, LDA and QDA show the same pattern of bias toward the larger group.

Of particular importance is that FMM and LDA/QDA tend to misclassify in different directions. When group sizes are equal, approximately equal numbers of cases are misclassified in either direction for FMM and LDA/QDA. However, when group sizes are unequal, FMM misclassifies in favor of the smaller group whereas LDA and QDA misclassify 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 LDA/QDA misclassify fewer cases from the larger group. As sample size and effect size are increased, these numbers are decreased even further.

Study 2

Study 2 followed the same procedure as Study 1 except data were generated to simulate nonrandom 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 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. Data were generated to the same conditions as in Study 1 (see Table 1) except data misclassification was generated in a nonrandom fashion. To achieve this type of misclassification, a random number was generated for each case and compared the cumulative probability of the point multiplied by a scalar that changed based on the desired proportion of misclassified cases. The cumulative probability was chosen because this quantity will be small when points are on the low end of the distribution (far from the overlap) and will be large when points are on the high end of the distribution (closer to the overlap) thus giving each point a number representing its relative standing in the distribution. The cumulative probability was used to calculate the percentage misclassified for the distribution with the smaller mean and the reverse cumulative probability was used to calculate the percentage misclassified for the distribution with the larger mean. In doing so, data far from its mean will receive low misclassification weights and data nearer the overlap of the distributions will receive higher misclassification weights. If the randomly generated number is less than the scaled cumulative probability, the point is misclassified to the other distribution. When this procedure is used without a scalar (or in other words, a scalar of 1 is used) approximately 50% of the cases will be misclassified on every iteration. Multiplying by an appropriate scalar, k, changes all of the probabilities by the same amount, thus making data either more or less likely for data to be misclassified. The values of k necessary to produce 100%, 90%, 80%, and 70% correctly classified data were determined through a proof using an application of single variable calculus. The formula for determining k by this method is as follows: percentage misclassified cases $= k/2$ (see Appendix A for proof).

Analyses. As in Study 1, FMM, LDA, and QDA were performed and the result was compared to the known classes. Again, the smaller group and larger group misclassification rates were calculated.

Results

Overall percentage correct for Study 2 is presented graphically in Figure 2. Again to save space, only results from the smallest sample size are shown (full set of results are available from the authors on request because the results are consistent with the small sample results.). In general, the results using nonrandom 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 (Table 3). As before, the initial misclassification does not affect the FMM solutions. Also, as expected, there is a decrease in LDA and QDA classification accuracy as the percentage of misclassified training data in the sample is increased.

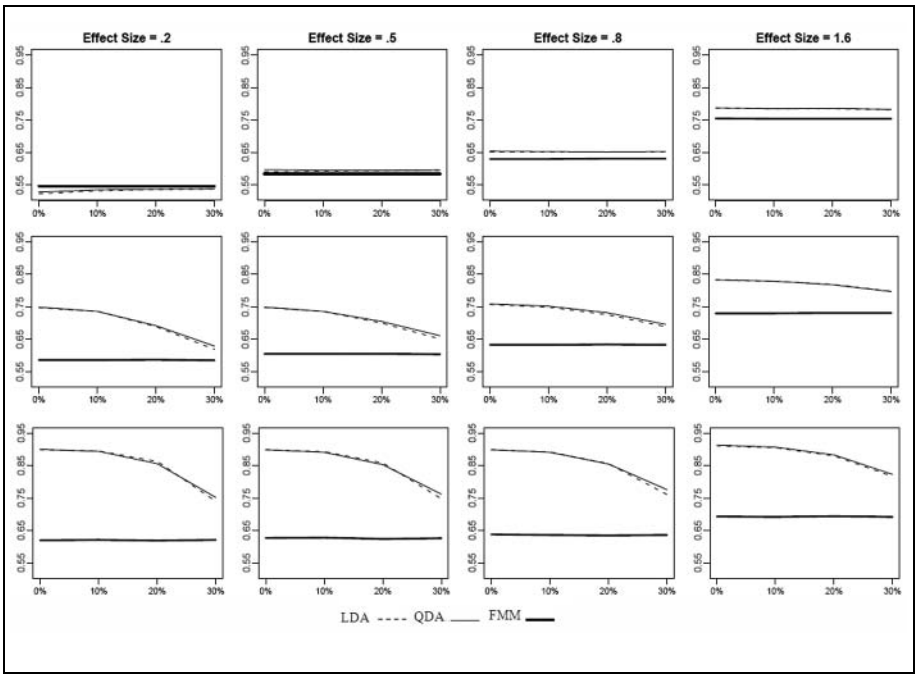


Figure 2. Percentage correct for Study 2 ($N = 100$)
Note: The *Y* axis presents the percentage of misclassified data. The *X* axis is the overall percentage correct achieved by the models. LDA = linear discriminant function analysis (dashed line); QDA = quadratic discriminant function analysis (thin solid line); FMM = finite mixture model (thick solid line).

Again mirroring the results from Study 1, we find LDA slightly outperforms QDA in all conditions. The difference between LDA and QDA is increased as the percentage of misclassified data is increased. The difference between LDA and QDA is vastly reduced, however, when sample size is increased. Again, similar to Study 1, although a difference between LDA and QDA exists, it is only marginal when comparing LDA and QDA to FMM.

Overall, LDA and QDA resulted in higher percentages of correctly classified data than did FMM. However, for the 50:50 condition at the lower effect sizes, FMM provided marginally higher percentage correct classification than did LDA and QDA. Again replicating the results of Study 1, the same pattern regarding the increase in classification due to sample size ratio emerges. As can be seen in Figure 2, for LDA and QDA, the increase in accuracy due to sample size ratio is quite large. For FMM, however, we see an increase in classification accuracy due to sample size ratio at the smaller effect sizes. As effect size increases the classification accuracy becomes more even across sample size ratios, and at the high effect sizes we see a decrease in classification accuracy due to sample size ratio. It should be noted that, we see exactly the same patterns for FMM in Study 1 and in Study 2 because FMM

Table 3. Small and Large Group Misclassification Percentages for Nonrandom Condition $N = 100$

			Percentage Misclassified											
			0%			10%			20%			30%		
		A:B	FMM	LDA	QDA	FMM	LDA	QDA	FMM	LDA	QDA	FMM	LDA	QDA
Small group misclassification rate	$\delta = 0.2$	10:90	0.060	0.100	0.100	0.060	0.099	0.099	0.060	0.092	0.093	0.060	0.075	0.074
		25:75	0.146	0.249	0.248	0.146	0.237	0.238	0.145	0.205	0.205	0.145	0.164	0.159
		50:50	0.227	0.236	0.238	0.227	0.234	0.235	0.228	0.231	0.232	0.227	0.231	0.232
	$\delta = 0.5$	10:90	0.056	0.100	0.099	0.056	0.096	0.097	0.055	0.085	0.088	0.055	0.065	0.064
		25:75	0.129	0.238	0.238	0.128	0.213	0.216	0.129	0.176	0.175	0.129	0.139	0.134
		50:50	0.208	0.202	0.205	0.206	0.202	0.203	0.208	0.203	0.203	0.209	0.202	0.202
	$\delta = 0.8$	10:90	0.050	0.097	0.098	0.049	0.091	0.093	0.049	0.076	0.079	0.049	0.055	0.054
		25:75	0.109	0.210	0.212	0.109	0.180	0.181	0.109	0.145	0.143	0.109	0.115	0.112
		50:50	0.186	0.173	0.174	0.186	0.173	0.174	0.183	0.173	0.174	0.184	0.173	0.173
	$\delta = 1.6$	10:90	0.032	0.073	0.074	0.031	0.063	0.064	0.032	0.046	0.048	0.032	0.028	0.030
		25:75	0.064	0.115	0.114	0.064	0.099	0.100	0.065	0.081	0.082	0.065	0.063	0.064
		50:50	0.121	0.107	0.107	0.123	0.107	0.107	0.124	0.107	0.108	0.123	0.108	0.108
Large group misclassification rate	$\delta = 0.2$	10:90	0.320	0.000	0.001	0.319	0.007	0.006	0.322	0.053	0.044	0.319	0.173	0.185
		25:75	0.268	0.003	0.005	0.267	0.027	0.026	0.270	0.103	0.106	0.271	0.207	0.222
		50:50	0.227	0.236	0.239	0.227	0.231	0.233	0.227	0.232	0.232	0.227	0.231	0.231
	$\delta = 0.5$	10:90	0.318	0.001	0.002	0.316	0.012	0.009	0.320	0.062	0.054	0.318	0.173	0.188
		25:75	0.266	0.013	0.014	0.267	0.050	0.048	0.267	0.119	0.125	0.266	0.200	0.214
		50:50	0.208	0.201	0.205	0.210	0.203	0.204	0.209	0.202	0.203	0.207	0.202	0.202
	$\delta = 0.8$	10:90	0.312	0.003	0.004	0.314	0.018	0.015	0.316	0.069	0.064	0.314	0.170	0.184
		25:75	0.259	0.031	0.031	0.257	0.068	0.070	0.258	0.123	0.131	0.257	0.189	0.199
		50:50	0.184	0.173	0.174	0.184	0.173	0.174	0.186	0.174	0.175	0.185	0.174	0.174
	$\delta = 1.6$	10:90	0.276	0.014	0.014	0.277	0.030	0.030	0.274	0.071	0.072	0.276	0.149	0.154
		25:75	0.206	0.052	0.053	0.205	0.071	0.071	0.205	0.101	0.099	0.206	0.140	0.138
		50:50	0.124	0.106	0.106	0.123	0.107	0.107	0.122	0.107	0.107	0.123	0.109	0.109

LDA = linear discriminant function analysis; QDA = quadratic discriminant function analysis; FMM = finite mixture model.

does 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. It should also be noted that, although the pattern exists, the classification differences between sample size ratios are still quite small, and only seem to make much of a negative impact at very large effect sizes (e.g., a standardized mean difference of 1.6). Thus, these results may not show cause for concern.

When looking at the direction of misclassification (shown in Table 3) a pattern similar to that in Study 1 emerges. Recall that in Study 1, when data were misclassified at random it was observed that FMM and LDA/QDA misclassify cases in opposite directions: FMM tends to misclassify in favor of the smaller group whereas LDA and QDA tend to misclassify in favor of the larger group. When data were misclassified systematically, in general we see the same pattern. However, in the 30% incorrect training data condition and the highest effect sizes of the 20% incorrect condition the pattern changes. The FMM pattern stays the same, but for LDA and QDA, the direction of misclassification reverses: LDA and QDA begin to misclassify in favor of the larger group instead of the smaller group. This effect is more extreme as effect size and discrepancy in group size are increased.

Comparing the random and nonrandom conditions, we can see that the random condition had a larger impact on classification accuracy of the LDA and QDA 50:50 sample size ratio than did the nonrandom condition. However, overall the 50:50 condition appears to be affected the least by misclassified data. The nonrandom condition had a stronger effect on classification Sample size ratio also did not seem to make as strong an impact in the nonrandom condition, though this is likely due to the stronger effect of the misclassified data lowering the accuracy of LDA and QDA.

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. Consistent with previous results increased sample size and effect size lead to increases in overall classification accuracy. Increased discrepancy in group sizes leads to increases in classification accuracy for LDA and QDA, but can actually decrease classification accuracy for large effect sizes for FMM. It is interesting that the two procedures assuming unequal variances appeared to be at a disadvantage when population variances were held equal. Furthermore, this study shows that the assumption of unequal variances may actually be a hindrance when variances are equal in the population.

This study shows that, consistent with Lachenbruch (1979), initial misclassification of groups done at random makes less of an impact on discriminant function methods than misclassification done in a nonrandom fashion when group sizes are equal. However, the impact of group size was not previously studied. Although Lachenbruch's

(1979) findings hold for equal group sizes, as group size becomes unequal, the impact of non-random misclassification becomes more important. Our study showed that for the extreme cases of misclassification (especially for large effect sizes) in sample size discrepant conditions, nonrandom misclassified data causes the direction of misclassification to reverse. It seems as though misclassified data does not actually cause LDA and QDA to reverse its direction of misclassification, but rather, the extreme amount of misclassified data causes so many classification errors that the proportions classified do not accurately represent how the data are being classified by the model. Neither type of misclassification affected the accuracy of FMM since initial classification is not part of the FMM.

In comparing accuracy between FMM, LDA, and QDA, LDA and QDA display 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. LDA showed slightly higher classification accuracy than QDA, especially as misclassified data was increased, replicating the previous findings of Lachenbruch (1979).

One of the most important findings of the study is the direction of misclassification for finite mixture modeling and DFA. Even though LDA and QDA achieve overall higher levels of classification accuracy in the majority of conditions and misclassify very little of the larger group, when group sizes are discrepant FMM better captures the smaller group. These results (found in Study 1) were largely replicated in Study 2 with nonrandom misclassified data (with the exception of LDA and QDA 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 the use of each technique is appropriate. Although some have documented the direction of misclassification bias for DFA (Breckenridge, 2000; Lei & Koehly, 2003), the comparison with FMM and implications for practical use are new.

Discussion

This article 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 FMM does not use in any way initial knowledge of group status. For DFA techniques, 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 nonrandom fashion.

These results have practical implications for the decision of when to use each technique. Because of its demonstrated high levels of accuracy, DFA may be the method of choice when the researcher is most interested in recovering the highest percentage correct (see Rausch & Kelley, 2009, in situations where DFA is appropriate but nonnormality may be present). However, as the results of both Study 1 and Study 2

indicate, there are times when FMM may provide a better alternative. In situations where the group sizes are approximately equal and expected effect size is low, FMM has demonstrated higher levels of classification accuracy. In situations where group sizes are unequal, although it demonstrates lower classification accuracy overall, FMM captures higher percentages 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 comprise approximately 3% to 10% of the overall student population (Hallahan, Keller, Martinez, Gelman, & Fan, 2007), and thus (when samples are representative of the population) we would see a sample size ratio of likely, at most, 10:90. An argument could be made that it is more desirable to err in the direction of initially misclassifying more students for screening purposes as having learning difficulties who do not, than to overlook students who truly have learning difficulties. This is a situation where FMM might provide a more desirable approach. Although the mixture model is likely to misclassify more nondisabled individuals as learning disabled than the DFA, the majority of truly disabled individuals would be identified.

Furthermore, 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 DFA. The choice of linear discriminant analysis versus quadratic discriminant analysis is also pertinent. Our results and the work of Lachenbruch (1979) suggest that when misclassified data is introduced into training data, the QDA classifies less effectively than the LDA. However, it is important to point out that this study was limited to holding population variances equal. It is unknown how unequal variances would affect LDA and QDA when misclassified data is introduced.

The robustness of DFA to random misclassification in training data comes as welcome news. However, it is somewhat worrisome that nonrandom misclassified data poses such a threat to classification accuracy in DFA, especially because it is more realistic in many situations. Although random misclassifications can happen, nonrandom misclassification because of implementation of cut score schemes seems far more likely to occur in practice. Even more worrisome, perhaps, is that researchers usually do not know the degree of misclassification. Thus, it is important to take possible effects of misclassification into account when interpreting results of a DFA.

Appendix A

Decision Rule for Nonrandom Misclassification

Misclassify point x if a uniform random variable between 0 and 1 is less than k times the cumulative probability of the point x. Or, more formally

$$\text{Given } x, \text{misclassify } x \text{ if } U < kF(x)$$

where U is a uniform random variable on the interval from 0 to 1, $F(x)$ is the cumulative distribution function, and k ($0 \leq k \leq 1$) is the scalar we are looking for to control the percentage of misclassified cases.

Thus defined, the total proportion of misclassified points can be defined as

$$\int_{-\infty}^{\infty} 1_{(U<kF(x))} f(x)dx,$$

where $f(x)$ is the density function corresponding to the cumulative distribution function $F(x)$. To find the expected total probability of misclassification (with respect to the random variable $U_{[0,1]}$) we use the rule

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

to get

$$E_u\left[\int_{-\infty}^{\infty} 1_{(U<kF(x))} f(x)dx\right] = \int_0^1 \left[\int_{-\infty}^{\infty} 1_{(U<kF(x))} f(x)dx\right] 1$$

$$du = \int_{-\infty}^{\infty} \left[\int_0^1 1_{(u<kF(x))} du\right] f(x)dx = \int_{-\infty}^{\infty} kF(x)f(x)dx.$$

Now, to evaluate the integral, we use the following result:

$$\int kF(x)f(x)dx = k\int F(x)f(x)dx = k\int vdv = \frac{k}{2}v^2 = \frac{k}{2}F(x)^2$$

Thus,

$$\int_{-\infty}^{\infty} kF(x)f(x)dx = \frac{k}{2}F(x)^2|_{-\infty}^{\infty} = \frac{k}{2}.$$

Appendix B

R Code Details

Data generation in R was achieved by using the R function rnorm() to create two classes of data with specified means and standard deviations (see Table 1). The distribution with the smaller mean was labeled "Distribution A,'' and the distribution with the larger mean was labeled "Distribution B.'' After data generation, the LDA, QDA, and FMM were performed on the misclassified data. The discriminant function analyses were performed using the lda() function for the LDA and qda() for the QDA, both located in the MASS (Venables & Ripley, 2002) R package. For LDA and QDA, the default settings for the analysis were used (prior probabilities estimated from the data, LDA assumed equal variances, QDA assumed unequal variances). The FMM was performed using the Mclust (data, G = 2, modelnames = c("V'')) function located in the mclust R package (Fraley & Raftery, 2002), with unequal variances (modelnames = c("V'')) and two groups (G = 2) 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.

Acknowledgments

The authors would like to thank Michael P. Trelinski and Elizabeth Housworth for insight on previous versions of this article.

Declaration of Conflicting Interests

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

Funding

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

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 2 x 2 tables. *Biometrics, 10*, 478-486.

Chhikara, R. S., & McKeon, J. (1984). Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association, 79*, 899-906.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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.

Flury, B. (1997). *A first course in multivariate statistics*. 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.

Grayson, D. A. (1987). Statistical diagnosis and the influence of diagnostic error. *Biometrics, 43*, 975-984.

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer-Verlag.

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

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363-385.

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

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.

Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. *Technometrics, 16*, 419-424.

Lachenbruch, P. A. (1979). Note on initial misclassification effects on the quadratic discriminant function. *Technometrics, 21*, 129-132.

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

Lei, P.-W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *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. (1972). Asymptotic results for discriminant analysis when initial samples are misclassified. *Technometrics, 14*, 415-422.

Milligan, G. W., Soon, S. C., & Sokol, L. M. (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*, 40-47.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rausch, J. R., & Kelley, K. (2009). A Comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods, 41*, 85-98.

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* (4th ed.). New York: Springer Science + Business Media.

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