ED 476 179 TM 034 902

AUTHOR	So, Tak-Shing Harry; Peng, Chao-Ying Joanne
TITLE	Comparison of K-Means Clustering with Linear Probability Model, Linear Discriminant Function, and Logistic Regression for Predicting Two-Group Membership.
PUB DATE	2002-04-00
NOTE	192p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC08 Plus Postage.
DESCRIPTORS	*Cluster Grouping; *Group Membership; *Prediction; Probability; Research Methodology; Simulation
IDENTIFIERS	*Mean (Statistics)

ABSTRACT

This study compared the accuracy of predicting two-group membership obtained from K-means clustering with those derived from linear probability modeling, linear discriminant function, and logistic regression under various data properties. Multivariate normally distributed populations were simulated based on combinations of population proportions, equality of covariance matrices, and group separation. The four statistical methods were applied to training samples drawn based on combinations of sample representativeness and sample size. Error rates were calculated based on the cross-validation results on test samples. The findings revealed that, depending on the data pattern, K-means clustering was a viable alternative when the accuracy of predicting the membership of the smaller population was the main objective. (Contains 25 figures, 32 tables, and 27 references.) (Author/SLD)

Comparisons of K-means Clustering with Linear Probability Model, Linear Discriminant Function, and Logistic Regression for Predicting Two-group Membership

Tak-Shing Harry So

Chao-Ying Joanne Peng

Indiana University-Bloomington

KEY WORDS: prediction, two-group membership, k-means clustering, linear probability model, linear discriminant function, logistic regression

Author Note

Send all correspondence to Chao-Ying Joanne Peng, Department of Counseling and Educational Psychology, School of Education, Room 4050, 201 N. Rose Ave., Indiana University, Bloomington, IN 47405-1006, tel: (812) 856-8337, fax: (812) 856-8333 or (812) 856-8400, email: peng@indiana.edu.

Paper presented at the 2003 Annual Meeting of American Educational Research Association, Chicago, IL: April 21-25, 2003.

Abstract

This study compares the accuracy of predicting two-group membership obtained from K-means clustering with those derived from linear probability modeling, linear discriminant function, and logistic regression under various data properties. Multivariate normally distributed populations were simulated based on combinations of population proportions, equality of covariance matrices, and group separation. The four statistical methods were applied to training samples were drawn based on combinations of sample representativeness and sample size. Error rates were calculated based on the cross-validation results on test samples. The findings revealed that, depending on the data pattern, K-means clustering was a viable alternative when the accuracy of predicting the membership of the smaller population was the main objective.

Comparisons of K-means Clustering with Linear Probability Model, Linear Discriminant Function, and Logistic Regression for Predicting Two-group Membership

For predictive studies in which the outcome variable is continuous, the ordinary least square (OLS) regression modeling is the most popular technique used in educational research (Elmore & Woehlke, 1996). However, when modeling dichotomous (or binary) outcome variables, alternative statistical techniques are available. These include linear discriminant function and logistic regression modeling (Cleary & Angel, 1984; Fraser, Jensen, Kiefer, & Popuang, 1994). Linear discriminant function (LDF) and logistic regression (LR) are considered viable alternatives to OLS regression for modeling dichotomous outcome variables (Long, 1997; Ryan, 1997; Tabachnick & Fidell, 2001; Yarnold, Hart, & Soltysik, 1994).

When the OLS regression model is applied to outcome variables, it is referred to as a linear probability model (LPM). The LPM is considered less suitable, theoretically, than either LR or LDF for prediction or classification (Long, 1997; Rice, 1994). However, the main advantage of LPM is that it is easily interpreted. And OLS regression is often taught in statistic courses required by Ph.D. programs (Aiken, West, Sechrest & Reno, 1990). Studies utilizing this technique to predict dichotomous outcomes are still found in the field of education (e.g., Grubb & Tuma, 1991; Kallio, 1995).

K-means clustering (KM) has seldom been employed in predictive studies. Indirect evidence supports the proposition that KM may yield better prediction or classification results than either LDF or LR. Wilson and Hardgrave (1995) compared the ability of a neural network technique (i.e., the back propagation training algorithm) with traditional methods, such as LDF or LR, for predicting the academic success of MBA students. Their result revealed that the neural network models performed at least as well as discriminant analysis or logistic regression. Balakrishnam, Cooper, Jacob and Lewis (1994) compared neural network techniques (i.e., the

Kohonen algorithm) with unsupervised learning with KM for classification. They concluded that KM outperformed the Kohonen algorithm in cluster recovery. Using a data set from a cancer research project, So (2002) examined the classification accuracy of KM, LDF, and LR and found the performance of KM to be superior to that of either LDF or LR. There has not been research documented in the literature that compares the predictive accuracy of KM with that of LDF, LR, or LPM in a two-group classification under various data properties. Thus, the current study seeks to fill this void by manipulating five data properties in simulated data sets and systematically examines the accuracy of predicting two-group membership by KM, LDF, LR, and LPM. The remainder of this paper is divided into eight sections: (1) Research Design, (2) Method, (3) Model Fitting, (4) Data Analyses, (5) Results of Group 1 Error Rates, (6) Results of Group 2 Error Rates, (7) Results of Total Error Rates, and (8) Implications for Educational Researchers.

1. Research Design

Five factors regarding data properties were manipulated including (1) population proportions (3 levels), (2) equality of covariance matrices (3 levels), (3) group separations (3 levels), (4) sample representativeness (3 levels), and (5) sample size (2 levels). The first three factors were related to features of the underlying population while the latter two were related to features of samples. Multivariate normal distributions were assumed for both populations.

Data Patterns

Two data patterns were utilized in order to assess the generalizability of the results (Tables 1 and 2). Data Pattern I had three variables while Data Pattern II had eight. Furthermore, the two data patterns differed in means as well as in variance-covariance structures. Both were previously used in Fan and Wang (1999) for comparing LDF and LR in the two-group classification problem.

Five Factors

Three levels under Factor 1 (population proportions) were 0.5:0.5, 0.25:0.75, and 0.1:0.9. Factor 2 (the equality of covariance matrices) had three levels: (a) the equal condition in which the two populations had equal covariance matrices, (b) the first unequal condition in which Population #1 had smaller covariances (one-fourth of the size of Population #2's covariances), and (c) the second unequal condition in which Population #1 had larger covariances (four times of the size of Population#2's covariances). The larger covariance matrix was set to be four times the smaller covariance matrix, following the studies of Fan and Wang (1999) and Lei and Koehly (2000). The ratio of 4 to 1 (or 1 to 4) reflected a moderate degree of variance heterogeneity, according to Hess, Olejnik, and Huberty (2001). The group separation factor (Factor 3) was quantified in terms of the Mahalanobis distance (d^2) between the two population means. It had three levels: 6.709, 2.236, and 0.745 for Data Pattern I, or 6.785, 2.262, and 0.754 for Data Pattern II. The first two levels were considered a large separation (Stevens, 1996; Meshbane & Morris, 1996) while the third level was considered a moderate separation, according to Huberty, Wisenbaker, and Smith (1987).

Factor 4 (sample representativeness) had three levels: (i) the modeled group was 20% over-sampled, (ii) sample proportions equaled to the population proportions, and (iii) the modeled group was 20% under-sampled. These levels were chosen so as to investigate the effect of sample representativeness and prior probabilities on prediction. For example, if the two population proportions were 0.1 and 0.9, the two sample proportions would be 0.12 (= 0.1+0.1×0.2) and 0.88 (=1−0.12), respectively, under the "20% over-sampled" condition.

Two sample sizes (200 or 400) were two levels of the fifth factor manipulated in this study. Samples of either size were randomly drawn from the two underlying populations. For example, if the population proportions were 0.1 and 0.9, a sample of 200 would consist of 24 randomly selected observations from Population #1 and 176 from Population #2 under the "20% over-sampled" condition (i.e., sample proportions of 0.12 to 0.88). These two sample sizes were considered moderately large by Fan and Wang (1999).

A fully crossed factorial design with 3×3×3×3×2 =162 combinations was adopted for the present study. For each combination, 200 samples were simulated from multivariate normal distributions according to both data patterns. Thus, a total of 64,800 [=(3×3×3×3×2×200)×2] samples were simulated with overlapping observations between the two groups. Each sample was analyzed by four statistical methods and error rates in prediction were noted.

2. Method

Multivariate Normal Populations

For each data pattern, multivariate normal distributions were simulated by the matrix decomposition method with an appropriate linear transformation (Mooney, 1997). The data simulation procedure was as follows:

1. Generate a $(n \times k)$ data matrix of values from normal distribution with mean and standard deviation of 0 and 1, respectively, where *n* is the number of observations and *k* is the number of variables.
2. Compute the Choleski decomposition, which is equivalent to the square root of the $(k \times k)$ correlation matrix specified by Data Pattern I or II (Tables 1 and 2).
3. Multiply the data matrix from (1) above with the Choleski decomposition from (2) to ensure that the correlation structure from Data Pattern I or II is built into the data matrix from (1).

4. Multiply the data matrix from (3) with the ($n \times k$) standard deviation matrix of Data Pattern I or II cellwise.

5. Add the ($n \times k$) mean matrix of Data Pattern I or II to the data matrix derived in (4).

The SAS® macro program for completing steps (1) to (5) may be obtained from the first author. The pseudo random number generator, the IML procedure, and the RANNOR function of SAS® Version 8.2, installed on an IBM RS/6000 SP machine with AIX 4.3 operating system, were employed in the execution of the SAS® macro program to simulate data matrices for Populations #1 and #2.

Population Size

The combined population size is set at 50,000 observations in order to control for the exact population proportions for the two groups. For example, under the 0.1:0.9 population proportions condition, 5,000 observations constituted Population #1 and 45,000 observations constituted Population #2. The ratio of population to sample was higher than 20 times, the criterion used in Fan and Wang (1999), or 15,000 observations used in Lei and Koehly (2000), to ensure that the populations simulated were sufficiently stable in terms of correlations among variables included in Data Pattern I or II.

3. Model Fitting

Once the two populations were simulated, a training sample was drawn first. The training sample was fitted with K-means clustering (KM), logistic regression model (LR), linear discriminant function (LDF), and linear probability model (LPM). For LR, LDF, and LPM, the probability of being from Population #1 was modeled.

Each training sample was first fitted with LR by the SAS® LOGISTIC procedure to determine the overlapping configuration of the training sample. When a complete separation or quasi-complete separation configuration was detected, a new training sample was drawn from the 50,000 observations simulated. The new training sample was once again examined for data separation configuration. This process continued until a training sample with overlapping configuration was obtained. At this point, the training sample was considered suitable for fitting all four statistical methods. A test sample, with size equal to that of the training sample, was subsequently drawn randomly from the remaining observations in the populations in order to compute the error rate.

Four SAS® procedures, FASTCLUS, REG, DISCRIM, and LOGISTIC were employed to carry out KM, LPM, LDF, and LR on the training sample. Statistical models derived from the training sample were subsequently cross-validated on the test sample. Error rates incurred in cross validations in predicting membership in Group 1, Group 2, and both groups (i.e., the overall) were tracked and used in further analyses.

4. Data Analysis

The performance of KM, LPM, LDF, and LR in predicting two-group membership under various data conditions was examined in a split-plot factorial (i.e., $\mathrm{SPF}_{33332\cdot 4}$) ANOVA design. In order to contrast these four statistical methods on common grounds, error rates derived from the KM model were compared to those from the LPM model with a 0.5 probability cut-off, from the LDF model with prior probabilities equal to sample proportions and a 0.5 probability cut-off, and from the LR model with a 0.5 probability cut-off. These four methods were considered levels of a "within-subjects" factor in the SPF design. The five factors, namely, (1) population proportion, (2) equality of covariance matrices, (3) group separation, (4) sample representativeness, and (5) sample size, were treated as "between-subjects" factors. The outcome variables were three error rates incurred in cross-validations. The Group 1 error rate was the proportion of observations in the test sample originated from Population #1 that were misclassified as belonging to Population #2. Similarly, the Group 2 error rate was the proportion of observations from Population #2 that were misclassified as belonging to Population #1. The Total error rate was the overall proportion of observations in the test sample that were misclassified.

A full $SPF_{33332 \cdot 4}$ ANOVA should have contained 63 main effects and interactions (i.e., 31 between-subjects effects and 32 within-subjects effects). In order to keep the results manageable and interpretable, a reduced ANOVA design was employed. This reduced ANOVA contained only 15 between-subjects effects (i.e., 5 main effects and 10 two-way interactions) and 16 within-subjects effects (i.e., 1 main effect, 5 two-way, and 10 three-way interactions). As a result, for each data pattern, three separate reduced ANOVAs were executed, one for each error rate. A total of six (i.e., three error rates by two data patterns) split-plot factorial ANOVAs were carried out.

The ANOVAs were performed by the SAS® GLM procedure. All the effects examined were considered as fixed effects. Because a large number of F tests (i.e., 31) were performed for each ANOVA, an alpha level of .0016 was employed in assessing the statistical significance of each effect. The overall alpha for each ANOVA was kept at a .05 level.

In addition to the ANOVA results, the eta squared and the partial omega squared (Maxwell, Camp, & Arvey, 1981) for each effect were computed. The eta squared (η^2) represents the proportion of sample total variance of the dependent measure explained by a particular effect. This index is defined as

$$\eta^2 = \frac{SS_{effect}}{SS_{total}},$$

where SS_{effect} is the effect sum of squares and SS_{total} is the total sum of squares. The partial omega squared ($\omega^2_{partial}$) is an index of strength of association between a factor (i.e., main or interaction effect) and the dependent measure with the effects of other factors removed. This index is suitable for a factorial design (Kirk, 1995). It was calculated according to the formula:

$$\omega^2_{partial} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$$

where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×2×4 (= 129,600) in the ANOVA design. Unlike the significance test of an F ratio, $\omega^2_{partial}$ is not affected by sample size. According to Cohen (1988), a $\omega^2_{partial}$ value between .06 and .14 indicates a moderate association, while a value of .14 or greater is a large association. In this study, effects that had at least a moderate association with the dependent measure were operationally defined as practically significant. Main effects were comprehensively examined regardless of their $\omega^2_{partial}$ values. Because of the large degrees of freedom associated with the two error terms in the SPF ANOVA, only interaction effects with a $\omega^2_{partial}$ value greater than .06 were examined in greater details in later sections.

5. Results of Group 1 Error Rates

As defined previously, Group 1 (G1) error rate was the proportion of observations in test sample originated from Population #1 that were misclassified as belonging to Population #2. The means and standard deviations of this error rates based on 200 replications for each combination of levels of the five factors for Data Pattern I are summarized in Table G1-1. Similarly, the means and standard deviations for Data Pattern II are presented in Table G1-2. The ANOVA results of G1 error rates for Data Pattern I are shown in Table G1-3. Out of the 31 main and interaction effects, 23 effects were statistically significant at the .0016 level. The ANOVA results of G1 error rates for Data Pattern II are presented in Table G1-4. Twenty-nine effects were statistically significant at the .0016 level for Data Pattern II.

Data Property Main Effects

Main effects of the five "between-subjects" factors on G1 error rates were statistically significant at the .0016 level for both data patterns. For Data Pattern I, four out of five factors had a $\omega^2_{partial}$ value larger than .06 (see Table G1-3). These four factors were Factor 1 (population proportion), Factor 2 (equality of covariance matrices), Factor 3 (group separation), and Factor 4 (sample representativeness). By the operational definition previously established for $\omega^2_{partial}$, these four factors were considered practically significant.

For Data Pattern II, only three of the five factors had a $\omega^2_{partial}$ value larger than .06 (see Table G1-4). These three factors were Factor 1 (population proportion), Factor 3 (group separation), and Factor 4 (sample representativeness).

Population Proportion

For Data Pattern I, Factor 1 (population proportion) explained 23.15% ($\eta^2 = .2315$) of the sample variance of G1 error rates. This factor had a strong association ($\omega^2_{partial} = .614$) with G1 error rates. The mean error rates for the three levels of population proportions were .593, .426, and .215. The Newman-Keuls procedure was performed to compare pairs of mean error rates. The Newman-Keuls procedure was selected on the basis of its excellent power and its capability of maintaining the nominal familywise type I error when the factor has only three levels. Results of the pairwise comparisons indicated that the three mean error rates were statistically significantly different from each other. When the population proportions were 0.1:0.9, the mean G1 error rate (.593) was the highest, while 0.5:0.5 split had the lowest mean error rate (.215).

For Data Pattern II, similar results were obtained. This factor explained 32.86% ($\eta^2 = .3286$) of the sample variance of G1 error rates. The $\omega^2_{partial}$ (.634) signified a strong association between this factor and the G1 error rate. Results from the Newman-Keuls procedure indicated that the mean error rate (.671) for the 0.1:0.9 population proportions was statistically significantly higher than that (.489) for the 0.25:0.75 condition. The mean error rate (.275) for the 0.5:0.5 split was the lowest and it was statistically significantly lower than the mean error rates for the other two conditions. Based on the results from both data patterns, it was concluded that the G1 error rate increased as the proportions of Population #1 and Population #2 further deviated from the 0.5:0.5 split.

Equality of Covariance Matrices

Factor 2 (equality of covariance matrices) had three levels: (a) equal covariance matrices, (b) Population #1 had smaller covariances that were one-fourth of the covariances of Population #2, and (c) Population #1 had larger covariances that were four times of the covariances of Population #2. For Data Pattern I, this factor explained 1.04% ($\eta^2 = .0104$) of the sample variance of G1 error rates. This factor exhibited a moderate association ($\omega^2_{partial} = .067$) with the G1 error rate. Results from the Newman-Keuls procedure indicated that the mean error rates for the three levels of this factor were statistically significantly different from each other. When Population #1 had smaller covariances, the mean error rate (.369) was the lowest; it was .416 for the equal covariance matrices condition and the highest (.449) when Population #1 had larger covariances.

For Data Pattern II, the equality of covariance matrices factor explained 0.51% ($\eta^2 = .0051$) of the sample variance of G1 error rates. This factor had a weak association ($\omega^2_{partial} = .026$) with the error rate. Results from the Newman-Keuls procedure indicated that the mean error rates for the three levels were statistically significantly different from each other. When Population #1 had smaller covariances, the mean error rate (.450) was the lowest; it was .490 for the equal covariance matrices condition and the highest (.495) when Population #1 had larger covariances. Results from both data patterns showed that, in general, sample observations from Population #1 were less likely to be misclassified when the covariances of Population #1 were smaller than that of Population #2.

Group Separation

Factor 3 (group separation) explained 24.87% ($\eta^2 = .2487$) of the sample variance of G1 error rates for Data Pattern I. A strong association ($\omega^2_{partial} = .631$) between this factor and G1 error rate was detected. Results from the Newman-Keuls procedure indicated that three mean error rates of group separation statistically significantly differed from each other. When the Mahalanobis distance (d^2) between the two populations' means was 6.709, the mean error rate was the lowest (.204). As d^2 decreased, the mean error rate increased. Under the condition of $d^2 = 2.236$, the mean error rate was .436; it was the highest (.594) when d^2 was 0.745.

For Data Pattern II, this factor accounted for 28.28% ($\eta^2 = .2828$) of the sample variance of G1 error rates. The association between group separation and the error rate was strong ($\omega^2_{partial} = .599$). Similar to results obtained from Data Pattern I, as d^2 decreased, the mean error rate increased for Data Pattern II. The mean error rate increased from .281 (when $d^2 = 6.785$) to .509 (when $d^2 = 2.262$) and further increased to .645 (when $d^2 = 0.754$).

Results from both data patterns led to the conclusion that the further the two populations separated, the lower was G1 error rate. These findings were expected because the less overlapping of the populations, the less likely that sample observations were misclassified.

Sample Representativeness

Factor 4 (sample representativeness) had three levels: (a) Group 1 (i.e., sample comprised of observations from Population #1) was 20% over-sampled, (b) sample proportions equal to population proportions, and (c) Group 1 was 20% under-sampled. For Data Pattern I, 1.46% ($\eta^2 = .0146$) of the sample variance of Group 1 error rates was explained by this factor. A moderate association ($\omega^2_{partial}$ = .091) was found between this factor and G1 error rates. Results from the Newman-Keuls procedure indicated that the three mean error rates sample representativeness differed statistically significantly from each other. The mean G1 error rate for the "20% over-sampled" condition was .365. It was slightly lower than the mean error rate (.409) under the "equal" condition and noticeably lower than the error rate (.460) under the "20% under-sampled" condition.

Similarly, for Data Pattern II, this factor explained 2.03% ($\eta^2 = .0203$) of the sample variance of G1 error rates. A $\omega^2_{partial}$ value of .0967 indicated a moderate association between this factor and the error rate. Results from the Newman-Keuls procedure indicated that the mean error rate for the "20% over-sampled" condition (.431) was statistically significantly lower than that under the "equal" condition (.475). The mean error rate of "20% under-sampled" condition (.529) was statistically significantly higher than those of the other two conditions. In general, results from both data patterns indicated that G1 error rate was low if Population #1 was over-sampled. When Population #1 was under-sampled, the error rate for G1 increased.

Sample size

Factor 5 (sample size) had two levels: 200 and 400. For Data Pattern I, this factor was not significant at $\alpha = .0016$. Less than 0.01% ($\eta^2 < .0001$) of the sample variance of G1 error rates was explained by this factor. The two levels of sample size were weakly associated with the error rate ($\omega^2_{partial} < .0001$). The mean G1 error rates for sample sizes of 200 and 400 were .412 and .411, respectively.

For Data Pattern II, the sample size factor was significant at $\alpha = .0016$. It accounted for less than 0.01% ($\eta^2 < .0001$) of the sample variance of the G1 error rate. The $\omega^2_{partial}$ value of .0001 indicated a virtually non-existent association. The mean error rates for sample sizes of 200 and 400 were .477 and .480, respectively.

By examining Tables G1-1 and G1-2, one notices that the standard deviations of G1 error rates for sample size of 400 are smaller than those for sample size of 200. In other words, with a larger sample size, one obtains a more efficient estimate of G1 error rates.

Two-way Interaction among Data Property Factors

For Data Pattern I, only 1 two-way interaction (i.e., population proportion by group separation) was judged to be practically significant, using $\omega^2_{partial} > .06$ as the criterion. For Data Pattern II, 2 two-way interactions among the five data property factors were considered practically significant. These two interactions were (a) population proportion by group separation interaction, and (b) equality of covariance matrices by group separation interaction. These are the only two-way interaction effects discussed here.

Population Proportion by Group Separation Interaction

This two-way interaction was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, it accounted for 2.44% ($\eta^2 = .0244$) of the sample variance of G1 error rates and showed a large association ($\omega^2_{partial}$ = .144) with the error rate. For Data Pattern II, this interaction explained 2.86% (η^2 = .0286) of the sample variance of G1 error rates and also exhibited a large association ($\omega^2_{partial}$ = .131). This interaction is graphically presented in Figure G1-1.

As shown in Figure G1-1, the mean G1 error rates for Data Pattern II were slightly higher than those for Data Pattern I. However, the interaction profiles for the two data patterns were similar. Regardless of the degree of group separation, the mean G1 error rate increased as the proportions of Population #1 and Population #2 deviated from the 0.5:0.5 split. The population proportion factor had a relatively small impact on the G1 error rate when the separation of the two populations was large (i.e., d^2 = 6.7). When the separation was smaller (i.e., d^2 = 2.2 or 0.7), the impact of population proportions on the error rate increased. The differences in mean G1 error rates among the three levels of group separation under the condition of 0.5:0.5 population proportions were relatively small, compared with the corresponding differences under the 0.1:0.9 condition.

Equality of Covariance Matrices by Group Separation Interaction

This two-way interaction explained 0.40% (η^2 = .0040) of variance in G1 error rates for Data Pattern I, and 1.43% (η^2 = .0143) for Data Pattern II. The $\omega^2_{partial}$ value (.027) for Data Pattern I indicated a weak association between the interaction effect and the error rate. However, for Data Pattern II, the interaction exhibited a moderate association ($\omega^2_{partial}$ = .070). This interaction is graphically presented in Figure G1-2.

As shown in Figure G1-2, the interaction profiles for the two data patterns were slightly different. For Data Pattern I, the association between the interaction and the error rate was weak.

The lines connecting the means of the three levels of equality of covariance matrices at each level of group separations are almost parallel to each other. Unequal covariance matrices had slightly stronger impact on G1 error rate when the group separation $d^2 = 6.7$ than when $d^2 = 2.2$ or 0.7. The differences among the mean error rates for the three levels of equality of covariance matrices under the condition of $d^2 = 2.2$ or 0.7 were minimal. The differences increased slightly when $d^2 = 6.7$.

For Data Pattern II, the impact of equality of covariance matrices on G1 error rate depended on the degree of group separation. When the group separation was large (i.e., $d^2 = 6.7$), the mean error rate for the condition in which Population #1 had smaller covariances was substantially lower than those for the other two conditions. However, when the group separation became smaller (i.e., $d^2 = 0.7$), the mean error rate for the condition in which Population #1 had smaller covariances was slightly higher than those found under the other two conditions.

Results from both data patterns indicated that the effect of equality of covariance matrices was the strongest when the two populations were well separated. When $d^2 = 6.7$, sample observations from Population #1 were less likely to be misclassified if the covariance matrices of Population #1 was smaller than that of Population #2.

Effects Concerning Statistical Methods

Did the use of different statistical methods result in significantly differences in the accuracy of predicting the membership of observations from Population #1? And did the five “between-subjects” factors related to data property have any joint impact with different statistical methods on G1 error rate? To answer these questions, the results of the “within-subjects” effects (i.e., the four statistical methods) were examined.

Main Effect of the Method Factor

The models from the four statistical methods (i.e., LPM, LDF, LR, and KM) were treated as levels of the "method" factor. This factor was statistically significant at $\alpha = .0016$. It accounted for 19.08% ($\eta^2 = .1908$) and 1.67% ($\eta^2 = .0167$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. $\omega^2_{partial}$ values of .866 and .205 for Data Patterns I and II, respectively, indicated a strong association between statistical methods and G1 error rates.

The mean G1 error rate for LPM was .545 for Data Pattern I, and .541 for Data Pattern II. The mean error rates for LDF were .464 and .462 for Data Patterns I and II, respectively. The mean error rate for LR was .462 for Data Pattern I and .458 for Data Pattern II. The mean error rates for KM were .175 and .453 for Data Patterns I and II, respectively. The Dunn-Šidák procedure was performed for pairwise comparisons among the four mean error rates. The selection of the Dunn-Šidák procedure was based on its excellent power and capability of exactly maintaining a small familywise type I error rate, such as .0016 (Kirk, 1995). Results from the Dunn-Šidák procedure indicated that the mean G1 error rates from the four methods were statistically significantly different from each other for both data patterns. LPM yielded the highest mean G1 error rate. The mean error rates for LDF were only slightly higher than those for LR even though the differences were statistically significant. The mean error rates for KM were the lowest.

Results indicated that LDF and LR performed similarly and both methods outperformed LPM. The performances of KM were not consistent for the two data patterns. KM either outperformed or performed as well as the other three methods.

Two-way Method by Data Property Interactions

For both data patterns, the method factor was found to be statistically significantly interacting with all five data property factors at $\alpha = .0016$. According to the criterion of $\omega^2_{partial} > .06$, the method by sample size interaction was the only interaction effect not considered practically significant. The other four interaction effects, considered practically significant, are discussed below.

Method by population proportion interaction. This two-way interaction accounted for 12.05% ($\eta^2 = .1205$) and 10.48% ($\eta^2 = .1048$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. The $\omega^2_{partial}$ values of .803 for Data Pattern I, and .618 for Data Pattern II represented a strong association between this interaction and the error rate. The interaction is presented graphically in Figure G1-3. The means and standard deviations are summarized in Table G1-5.

As shown in Figure G1-3 and Table G1-5, KM's performance was independent of the three conditions of population proportions. However, the mean error rates of KM for Data Pattern I were lower than those for Data Pattern II. For Data Pattern I, KM outperformed the other three methods. When the population proportions were 0.5:0.5, the mean G1 error rate for KM was slightly lower than those for LPM, LDF, or LR. When the population proportions approached the extreme condition (i.e., 0.1:0.9), the mean error rate for KM was substantially lower than those for the other three methods. For Data Pattern II, KM outperformed the other three methods only when the population proportions were 0.25:0.75 or 0.1:0.9. When the population proportions were 0.5:0.5, KM performed poorly, compared with the other three methods. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: when the population proportions were extreme (i.e., 0.1:0.9), KM was the best method; LPM was not the method of choice when the population proportions were extreme. When the population proportions were 0.25:0.75, KM remained to be a viable alternative method. When the population proportions were 0.5:0.5, the performances of LPM, LDF, and LR were identical; KM could perform as well as the other three methods.

Method by equality of covariance matrices interaction. This two-way interaction accounted for 2.17% ($\eta^2 = .0217$) and 0.44% ($\eta^2 = .0044$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. A strong association was found between this interaction and the error rate ($\omega^2_{partial} = .424$) for Data Pattern I. However, for Data Pattern II, the $\omega^2_{partial}$ value (.063) signified a moderate association. The means and standard deviations are summarized in Table G1-6. The interaction is presented graphically in Figure G1-4.

As shown in Figure G1-4, unequal covariance matrices conditions had relatively small impact on the differential performances of LPM, LDF, and LR. Yet, the mean G1 error rates of KM depended on the degree of equality of covariance matrices. The performances of LPM, LDF, and LR were similar for each data pattern. However, KM performed quite differently for both data patterns.

In Table G1-6, for Data Pattern I, KM outperformed the other three methods at all levels of the equality of covariance matrices. The performance of KM was exceptionally good when the covariances of Population #1 were one-fourth of those of Population #2. For Data Pattern II, KM outperformed LPM regardless of the condition of covariance matrices. However, it performed slightly better than LDF or LR when Population #1 had smaller covariances. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: all four methods performed better when the covariances of Population #1 were smaller than those of Population #2. KM was the best method under this condition. LPM was not the method of choice regardless of the equality of covariance matrices.

Method by group separation interaction. This two-way interaction accounted for 3.16% ($\eta^2 = .0316$) and 5.29% ($\eta^2 = .0529$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. The $\omega^2_{partial}$ index equaled .517 for Data Pattern I and .450 for Data Pattern II. A strong association between this interaction and the error rate was detected. This two-way interaction is graphically presented in Figure G1-5. The means and standard deviations of G1 error rate are summarized in Table G1-7.

As shown in Figure G1-5, the performances of the four methods in predicting the membership of Population #1 depended on the degree of separation between the two populations' means. The further the two population means separated, the lower was G1 error rate regardless which method was used. The performances of LPM, LDF, and LR were similar for each data pattern. However, KM performed differently for the two data patterns and also from LPM, LDF, and LR.

In Table G1-7, for Data Patten I, KM outperformed the other three methods regardless of the degree of group separation. The performance of KM was exceptionally good when the group separation was small (i.e., $d^2 = 2.2$ or 0.7). For Data Pattern II, when $d^2 = 0.7$, KM performed better than the other three methods. When $d^2 = 2.2$, the mean G1 error rate for KM was only slightly lower than those of the other three methods. When $d^2 = 6.7$, the performance of KM was the worst among the four methods. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: LPM or KM was a viable alternative only when the two populations were not well-separated (i.e., $d^2 = 2.2$ or 0.7); group separation had little impact on the relative efficiency of LR over LDF.

Method by sample representativeness interaction. This two-way interaction accounted for only 0.69% ($\eta^2 = .0069$) and 0.66% ($\eta^2 = .0066$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. However, for Data Pattern I, this interaction exhibited a strong association ($\omega^2_{partial} = .189$) with the error rate. For Data Pattern II, a moderate association ($\omega^2_{partial} = .092$) was detected. The interaction is graphically presented in Figure G1-6. The means and standard deviations of G1 error rate are summarized in Table G1-8.

As shown in Figure G1-6, the performance of the four methods depended on sample representativeness. The performances of LPM, LDF, and LR are similar for either data pattern but different from KM. The mean G1 error rates were the highest when Group 1 was "20% under-sampled;" the mean error rates were the lowest when Group 1 was "20% over-sampled." Although the performance of KM was different for the two data patterns, the mean error rates of KM were similar under the three conditions of sample representativeness.

In Table G1-8, the mean G1 error rates for LPM were slightly higher than those for LDF and LR regardless of the condition of sample representativeness. Meanwhile, LR performed slightly better than LDF in all three conditions of sample representativeness.

For Data Pattern I, KM outperformed the other three methods. When Group 1 was "20% over-sampled," the mean error rate was the highest. The mean error rate was the lowest when Group 1 was "20% under-sampled." However, the differences between the "equal" condition and the other two conditions were less than .01. For Data Pattern II, KM performed similarly regardless of the condition of sample representativeness. It outperformed the other three methods when Group 1 was either "equal" or "20% under-sampled." When Group 1 was "20% over-sampled," KM was not as good as LDF or LR, but better than LPM. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: KM was the method of choice especially when the sample representativeness was questionable. LPM was not a method of choice regardless of sample representativeness.

Three-way Method by Data Property Interactions

For Data Pattern I, 2 three-way interactions concerning statistical methods were considered practically significant (i.e., effects with $\omega^2_{partial} > .06$): (a) method by population proportion by group separation interaction, and (b) method by equality of covariance matrices by group separation interaction. However, for Data Pattern II, only the method by population proportion by group separation interaction was considered practically significant. In addition to the 2 three-way interactions, the method by population by equality of covariance matrices interaction is also included in this section. They are the only results discussed below.

Method by population proportion by group separation interaction. This three-way interaction explained 1.87% ($\eta^2 = .0187$) and 2.18% ($\eta^2 = .0218$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. This interaction exhibited a strong association ($\omega^2_{partial} = .388$ and .252 for Data Patterns I and II, respectively) with the error rate. The interaction is graphically presented in Figure G1-7. Separate plots of mean G1 error rates due to the population proportion by group separation interaction are presented in (a), (b), (c), and (d) for the four methods. The means and standard deviations for the interaction are summarized in Table G1-9.

As shown in Figure G1-7, the interaction profiles for LPM, LDF, and LR are similar for the two data patterns. The mean G1 error rates were the highest when the population proportions were 0.1:0.9 while those for the 0.5:0.5 condition were the lowest regardless of the degree of group separation. However, the impact of group separation on G1 error rate was intensified when the population proportion deviated from the 0.5:0.5 split.

Figure G1-7(d) illustrates the population proportion by group separation interaction under KM. The interaction profiles were different for the two data patterns. For Data Pattern I, the mean G1 error rates increased as the separation between the two populations decreased regardless of population proportions. The impact of population proportions on G1 error rate was consistent for the three levels of group separation. Unlike the other three methods, the mean error rates for KM were the lowest when the population proportions were 0.1:0.9 and the highest for the 0.5:0.5 condition. Compared with the other three methods, KM outperformed the others under all conditions except when the population proportions were 0.5:0.5 and $d^2 = 6.7$. For Data Pattern II, the mean G1 error rates increased as the separation between the two populations decreased regardless of population proportions. KM performed similarly regardless of population proportions when $d^2 = 2.2$ or 0.7. However, when $d^2 = 6.7$, KM performed the best under the 0.5:0.5 population proportions. Compared with the other three methods, KM outperformed the others under all conditions of population proportions when $d^2 = 2.2$ or 0.7. When $d^2 = 6.7$, KM performed poorly especially under the population proportions of 0.50:0.5 or 0.25:0.75. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results indicated that (a) LPM was not the method of choice in predicting the membership of Population #1 especially when the population proportions were extreme and the two populations were well-separated, (b) KM was a viable alternative when population proportions deviated from the 0.5:0.5 split, regardless of the degree of group separation.

Method by equality of covariance matrices by group separation interaction. This three-way interaction explained 0.63% ($\eta^2 = .0063$) and 0.18% ($\eta^2 = .0018$) of the sample variance of G1 error rates for Data Patterns I and II, respectively. For Data Pattern I, although this interaction accounted for less than 1% of the sample variance of the error rate, a strong association ($\omega^2_{partial} = .175$) between the interaction and the error rate was detected. For Data Pattern II, a weak association ($\omega^2_{partial} = .027$) was found. The means and standard deviations of G1 error rate for the three-way interaction are summarized in Table G1-10. The interaction is graphically presented in Figure G1-8. Separate plots of mean G1 error rates due to the equality of covariance matrices by group separation interaction are presented in (a), (b), (c), and (d) for the four methods.

As shown in Figure G1-8, the interaction profiles for LPM, LDF, and LR are similar for the two data patterns. The plots illustrate that the impact of equality of covariance matrices on G1 error rate depended on the degree of group separation. When the group separation was large (i.e., $d^2 = 6.7$), the mean error rate for the condition in which Population #1 had smaller covariances was substantially lower than those for the other two conditions. However, when the group separation became small (i.e., $d^2 = 2.2$ or 0.7), the mean error rate for the condition in which Population #1 had smaller covariances was slightly higher than those for the other two conditions.

In Figure G1-8(d), the mean G1 error rates are plotted for the equality of covariance matrices by group separation interaction under KM. The interaction profiles were different for the two data patterns. For Data Pattern I, the mean error rates increased as the degree of group separation decreased regardless of the condition of equality of covariance matrices. In Table G1-10, the mean error rates for the condition in which Population #1 had smaller covariances were consistently the lowest at each level of the group separation. For Data Pattern II, a similar conclusion was reached. However, the impact of the equality of covariance matrices factor on the error rate was small when $d^2 = 0.7$ or 2.2. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results indicated that (a) for LPM, LDF, and LR, a strong impact of heterogeneity of covariance matrices on predicting the membership of Population #1 was found only when the two populations were well-separated, and (b) KM was a viable alternative when the separation between the two populations' means were small (i.e., $d^2 = 2.2$ or 0.7) regardless of the equality of covariance matrices.

Method by population proportion by equality of covariance matrices interaction. This three-way interaction explained 0.15% ($\eta^2 = .0015$) of the sample variance of G1 error rates for both data patterns. This interaction exhibited a weak association ($\omega^2_{partial} = .048$ and .022 for Data Patterns I and II, respectively) with the error rate. This interaction is included here in order to contrast the results of this study with the findings from Fan and Wang (1999). The means and standard deviations of G1 error rate for the three-way interaction are summarized in Table G1-11.

In Table G1-11, the G1 error rates of LPM were higher than those of the other three methods when the population proportions deviated from the 0.5:0.5 split. LPM and LDF performed similarly when the population proportions were 0.5:0.5, regardless of the condition of equality of covariance matrices. The difference in performance of LDF and LR depended on the combinations of population proportions and equality of covariance matrices.

The performance of KM was different for the two data patterns. For Data Pattern I, KM outperformed the other three methods in all joint conditions except when population proportions were 0.5:0.5 and Population #1 had larger covariances. For Data Pattern II, KM performed better than the other three methods under four joint conditions: population proportions were either 0.1:0.9 or 0.25:0.75 and the two populations either had equal covariance matrices or Population #1 had smaller covariances. The direction of impact of heterogeneity of covariance matrices on the performances of KM was consistent for both data patterns regardless of population proportions. The condition in which Population #1 had smaller covariances exhibited a small but positive effect on G1 error rates, while the condition in which Population #1 had larger covariances had a small but negative effect.

Results indicated that (a) LPM was not the method of choice when population proportions deviated from 0.5:0.5 split, (b) the direction of impact of heterogeneity of covariance matrices on the performance of LPM, LDF, and LR depended on population proportions, (c) the direction of impact of heterogeneity of covariance matrices on the performance of KM was consistent regardless of population proportions, (d) selection of LDF or LR required the consideration of both population proportions and the heterogeneity of covariance matrices, and (e) if the G1 error rate was the main concern, KM was a viable method especially when population proportions deviated from 0.5:0.5 and Population #1 had smaller covariances.

6. Results of Group 2 Error Rates

Group 2 (G2) error rate was the proportion of observations in test sample originated from Population #2 that were misclassified as belonging to Population #1. The means and standard deviations of the error rates of 200 replications for each combination of levels of the five factors for Data Pattern I are summarized in Table G2-1. Similarly, the means and standard deviations for Data Pattern II are presented in Table G2-2. The results of ANOVA on G2 error rates for Data Pattern I are summarized in Table G2-3. Out of 31 main and interaction effects, 23 effects were statistically significant at the .0016 alpha level. The ANOVA results for G2 error rate for Data Pattern II are presented in Table G2-4. Twenty-eight effects were statistically significant at the .0016 level for Data Pattern II.

Data Property Main Effects

For Data Pattern I, four of the five "between-subjects" (i.e., data property) main effects on G2 error rate were statistically significant at the .0016 level. These four factors also had a $\omega^2_{partial}$ value larger than .06 (see Table G2-3). For Data Pattern II, all five main effects were statistically significant at $\alpha = .0016$. However, only the same four factors had a $\omega^2_{partial}$ value larger than .06 (see Table G2-4). These four factors were Factor 1 (population proportion), Factor 2 (equality of covariance matrices), Factor 3 (group separation), and Factor 4 (sample representatives). Thus, these four factors were considered practically significant.

Population Proportion

For Data Pattern I, Factor 1 (population proportion) explained 12.52% ($\eta^2 = .1252$) of the sample variance of G2 error rates. This factor demonstrated a strong association ($\omega^2_{partial} = .421$) with the error rate. The mean error rates for the three levels of population proportions were .096, .101, and .216. Results from the Newman-Keuls pairwise procedure indicated that the three mean error rates were statistically significantly different from each other. When population proportions were 0.1:0.9, the mean G2 error rate (.096) was the lowest, while the 0.5:0.5 condition had the highest mean error rate (.216).

Similarly, for Data Pattern II, this factor explained 9.76% ($\eta^2 = .0976$) of the sample variance of G2 error rates. The $\omega^2_{partial}$ index (.483) signified a strong association. Results from the Newman-Keuls procedure indicated that the mean error rate (.275) for the 0.5:0.5 population proportion condition was statistically significantly higher than that (.158) for the 0.25:0.75 condition. The mean error rate (.131) for the 0.1:0.9 condition was the lowest and it was statistically significantly lower than the means for the other two conditions. Based on the results from both data patterns, we concluded that G2 error rate decreased, in general, as the proportions of Population #1 and Population #2 deviated from the 0.5:0.5 split.

Equality of Covariance Matrices

Factor 2 (equality of covariance matrices) had three levels: (a) equal covariance matrices, (b) Population #1 had smaller covariances that were one-fourth of the covariances of Population #2, and (c) Population #1 had larger covariances that were four times of the covariances of Population #2. For Data Pattern I, this factor explained 5.75% ($\eta^2 = .0575$) of sample variance of G2 error rates. This factor exhibited a large association ($\omega^2_{partial}$ = .250) with the error rate. Results from the Newman-Keuls procedure indicated that the mean error rates were statistically significantly different from each other. Under the unequal condition in which Population #2 had larger covariances, the mean G2 error rate (.179) was the highest. The mean error rate was .147 for the equal covariance matrices condition. Under the unequal covariance condition in which Population #2 had smaller covariances, the mean error rate (.088) was the lowest.

Similar results were obtained for Factor 2 from Data Pattern II. This factor explained

1.86% (η^2 = .0186) of the sample variance of G2 error rates. This factor had a large association ($\omega^2_{partial}$ = .151) with the error rate. Results from the Newman-Keuls procedure indicated that the mean error rates for the three levels were statistically significantly different from each other. Under the unequal condition that Population #2 had larger covariances, the mean error rate was the highest (.217). The mean error rate was .196 for the condition of equal covariance matrices. The mean error rate was the lowest (.151) for the unequal condition in which Population #2 had smaller covariances. Results from both data patterns showed that observations from Population #2 were less likely to be misclassified when Population #2 had smaller covariances than those of Population #1.

Group Separation

Factor 3 (group separation) explained 9.27% (η^2 = .0927) of the sample variance of G2 error rates for Data Pattern I. A strong association ($\omega^2_{partial}$ = .350) between this factor and G2 error rates was detected. Results from the Newman-Keuls procedure indicated that the mean error rates of the three levels of group separation differed from each other. When the Mahalanobis distance (d^2) between the two populations' means was 6.709, the mean error rate was .075. As d^2 decreased, the mean error rate also increased. Under the condition of d^2 = 2.236, the mean error rate was .147. The mean error rate was .191 when d^2 = 0.745.

For Data Pattern II, this factor accounted for 2.69% (η^2 = .0269) of the sample variance of G2 error rates. The association between this factor and the error rate was strong ($\omega^2_{partial}$ = .205). Similar to the results obtained from Data Pattern I, as d^2 decreased, the mean G2 error rate increased for Data Pattern II. The mean error rate increased from .144 (when d^2 = 6.785) to .197 (when d^2 = 2.262) and further increased to .223 (when d^2 = 0.754). Results from both data patterns led to the conclusion that the further the two populations separated, the lower was G2 error rate. These findings were expected because the less overlapping between the two populations, the less likely that sample observations would be misclassified.

Sample Representativeness

Factor 4 (sample representativeness) had three levels: (a) Group 1 (i.e., sample comprised of observations from Population #1) was 20% over-sampled, (b) sample proportions equal to the population proportions, and (c) Group 1 was 20% under-sampled. When Group 1 was over-sampled, Group 2 was under-sampled. Similarly, when Group 1 was under-sampled, Group 2 was over-sampled. When the population proportions were 0.1:0.9, Group 2 was either 2.2% under-sampled or over-sampled. When the population proportions were 0.25:0.75, Group 2 was either 6.7% under-sampled or over-sampled. When the population proportions were 0.5:0.5, the sample representativeness condition of Group 2 was either 20% under-sampled or over-sampled.

This factor was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, only 1.82% ($\eta^2 = .0182$) of the sample variance of G2 error rate was explained by this factor. A moderate association ($\omega^2_{partial} = .096$) was found between this factor and the error rate. Results from the Newman-Keuls procedure indicated that the mean error rates of the three levels differed statistically significantly from each other. The mean G2 error rate for the "20% over-sampled" condition was .165 for Data Pattern I. It was slightly higher than the mean error rate under the "equal" condition (.135). The "20% under-sampled" condition had the lowest mean error rate (.114).

Similarly, for Data Pattern II, this factor explained only 1.27% ($\eta^2 = .0127$) of the sample variance of G2 error rate. The $\omega^2_{partial}$ value (.108) indicated a moderate association between this factor and the error rate. Results from the Newman-Keuls procedure indicated that the mean error rate for the “20% over-sampled” condition (.217) was statistically significantly higher than that under the “equal” condition (.186). The mean error rate of “20% under-sampled” condition (.162) was the lowest which was statistically significantly smaller than those of the other two conditions.

Results from both data patterns indicated that, in general, there was a tendency for G2 error rate to increase if Group 1 was over-sampled, hence, Group 2 was under-sampled. When Group 1 was under-sampled, therefore, Group 2 was over-sampled, the G2 error rate decreased.

Sample size

Factor 5 (sample size) had two levels: 200 and 400. For Data Pattern I, less than 0.01% ($\eta^2 < .0001$) of the sample variance of G2 error rate was explained by this factor. This factor was weakly associated with the error rate ($\omega^2_{partial} < .0001$). The mean error rates for both sample sizes were .138. Similar results were obtained from Data Pattern II. The sample size factor accounted for less than 0.01% ($\eta^2 < .0001$) of the sample variance of G2 error rate. $\omega^2_{partial}$ (.0002) indicated a virtually non-existent association. The mean error rates for sample sizes of 200 and 400 were .189 and .187, respectively.

By examining Tables G2-1 and G2-2, one notices that standard deviations of G2 error rates for sample size of 400 were smaller than those for sample size of 200. In other words, with a larger sample size, one obtains a more efficient estimate of G2 error rate.

Two-way Interaction Among Data Property Factors

Using the $\omega^2_{partial} > .06$ as the criterion for practical significance, 3 two-way interactions among the five factors were considered practically significant for both data patterns. These three interactions were (a) population proportion by equality of covariance matrices, (b) population proportion by group separation, and (c) population proportion by sample representativeness.

They are the only results discussed below.

Population Proportion by Equality of Covariance Matrices

This two-way interaction was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, it accounted for 9.27% ($\eta^2 = .0927$) of the sample variance of G2 error rate and had a large association ($\omega^2_{partial} = .350$) with the error rate. However, for Data Pattern II, this interaction explained 0.74% ($\eta^2 = .0074$) of the sample variance of the error rate and also exhibited a moderate association ($\omega^2_{partial} = .066$). This interaction is graphically presented in Figure G2-1.

As shown in Figure G2-1, similar interaction profiles were found for the two data patterns. When the population proportions were 0.1:0.9 or 0.25:0.75, the effect of equality of covariance matrices on the error rate was relatively small. However, when the population proportions were 0.5:0.5, the mean G2 error rate under the condition in which Population #2 had smaller covariances was considerably lower than those of the other two covariance matrix conditions.

Population Proportion by Group Separation Interaction

This two-way interaction was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, it accounted for 4.01% ($\eta^2 = .0401$) of the sample variance of G2 error rates and had a large association ($\omega^2_{partial} = .277$) with the error rate. For Data Pattern II, this factor explained 4.74% ($\eta^2 = .0474$) of the sample variance of G2 error rate and also exhibited a large association ($\omega^2_{partial} = .216$). This interaction is graphically presented in Figure G2-2.

As shown in Figure G2-2, the interaction profiles for the two data patterns were similar. The population proportion factor had a relatively small effect on G2 error rate when the separation of the two populations was large (i.e., $d^2 = 6.7$). However, when the separation was smaller (i.e., $d^2 = 2.2$ or 0.7), the impact of population proportion on G2 error rate increased. The differences of mean error rates among the three levels of group separation under 0.5:0.5 population proportions were large, compared with corresponding differences under either the 0.1:0.9 or the 0.25:0.75 condition.

Population Proportion by Sample Representativeness Interaction

This two-way interaction was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, it accounted for 2.44% ($\eta^2 = .0244$) of the sample variance of G2 error rate and had a moderate association ($\omega^2_{partial} = .124$) with the error rate. For Data Pattern II, this factor explained 1.33% ($\eta^2 = .0133$) of the sample variance of the error rate and also exhibited a moderate association ($\omega^2_{partial} = .0113$). This interaction is graphically presented in Figure G2-3.

As shown in Figure G2-3, similar interaction profiles were found for the two data patterns. G2 error rate increased if Group 1 was over-sampled, hence, Group 2 was under-sampled. When Group 1 was under-sampled, therefore, Group 2 was over-sampled, the G2 error rate decreased. The sample representativeness factor had a small impact on G2 error rates when the population proportions were 0.1:0.9. The impact of sample representativeness on the error rate increased as population proportions approached 0.5:0.5.

Effects Concerning Statistical Methods

The results of the "within-subjects" effects (i.e., four statistical methods) were examined in order to investigate: (a) the effect of using different statistical methods on G2 error rates, and (b) the joint effect of the five data property factors with statistical methods on G2 error rate.

Main Effect of Method Factor

The models from the four statistical methods (i.e., LPM, LDF, LR, and KM) were treated as levels of the "method" factor. This factor was statistically significant at $\alpha = .0016$. It accounted for 26.69% ($\eta^2 = .2669$) and 64.35% ($\eta^2 = .6435$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ values (.883 and .944 for Data Patterns I and II, respectively) indicated a strong association between statistical methods and G2 error rate.

The mean G2 error rate for LPM was .086 for Data Pattern I, and .089 for Data Pattern II. The mean error rates for LDF were .093 and .097 for Data Patterns I and II, respectively. The mean error rate for LR was .095 for Data Pattern I and .100 for Data Pattern II. The mean error rates for KM were .278 and .467 for Data Patterns I and II, respectively. Results from the Dunn-Šidák comparison procedure indicated that the mean G2 error rates from the four methods were statistically significantly different from each other for both data patterns. LPM yielded the lowest mean G2 error rate while KM the highest. The mean error rates for LDF were only slightly lower than those for LR even though the differences were statistically significant.

Results indicated that, in general, KM did not perform as well as the other three methods in predicting the membership of Population #2 for both data patterns. LPM outperformed LDF and LR whereas LDF and LR performed similarly.

Two-way Method by Data Property Interactions

The method factor was found statistically significantly interacting with all five data property factors at $\alpha = .0016$ for both data patterns. For Data Pattern I, the method by sample size interaction was the only interaction effect considered not to be practically significant according to the criterion of $\omega^2_{partial} > .06$. For Data Pattern II, only two data property factors (i.e., population proportions and sample representativeness) were found practically significantly interacting with the method factor. The interactions between the method factor and four data property factors (i.e., population proportion, equality of covariance matrices, group separation, and sample representativeness) are discussed in this section.

Method by population proportion interaction. This two-way interaction effect accounted for 16.64% ($\eta^2 = .1664$) and 5.84% ($\eta^2 = .0584$) of the sample variance of G2 error rates for Data Patterns I and II, respectively. The $\omega^2_{partial}$ values (.824 for Data Pattern I and .603 for Data Pattern II) indicated a strong association between this interaction and the error rate. The interaction is presented graphically in Figure G2-4. The means and standard deviations are summarized in Table G2-5.

As shown in Figure G2-4, the performance of the four methods depended on sample representativeness. The performance of KM was inconsistent across the two data patterns. In Table G2-5, for Data Pattern I, KM slightly outperformed the other three methods in predicting the membership of Population #2 only when the population proportions were 0.5:0.5. When population proportions were 0.25:0.75 or 0.1:0.9, G2 error rates of KM were substantially higher than those for the other three methods. For Data Pattern II, KM performed poorly, compared to the other three methods in all three population proportion conditions. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: when the population proportions deviated from 0.5:0.5, LPM was the best method. KM was not the method of choice when the population proportions were extreme. When the population proportions were 0.5:0.5, the performances of LPM, LDF, and LR were identical; KM could perform as well as the other three methods for Data Pattern I, but not for Data Pattern II.

Method by equality of covariance matrices interaction. This two-way interaction effect accounted for 3.31% ($\eta^2 = .0331$) and 0.18% ($\eta^2 = .0018$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.482) indicated a strong association between this interaction and the error rate for Data Pattern I. However, for Data Pattern II, $\omega^2_{partial}$ (.045) signified a weak association. The interaction is presented graphically in Figure G2-5. The means and standard deviations are summarized in Table G2-6.

As shown in Figure G2-5, unequal covariance matrices conditions had small impacts on the differential performances of LPM, LDF, and LR. Yet, the mean G2 error rates of KM depended on the degree of equality of covariance matrices. The performances of LPM, LDF, and LR were similar for either data pattern. However, KM performed quite differently for the two data patterns, also from LPM, LDF, and LR. All four methods performed better when the covariances of Population #2 were four times of those of Population #1.

In Table G2-6, for both data patterns, KM performed poorly compared to the other three methods regardless of the equality of covariance matrices. The G2 error rates of KM for Data Pattern I were lower than those of Data Pattern II. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: all four methods performed better when the covariances of Population #1 were larger than those of Population #2. LPM was uniformly the best method. KM was not the method of choice regardless of the condition of equality of covariance matrices.

Method by group separation interaction. This two-way interaction effect accounted for 2.56% ($\eta^2 = .0256$) and 0.06% ($\eta^2 = .0006$) of the sample variance of G2 error rate for Data

Patterns I and II, respectively. $\omega^2_{partial}$ (.419) indicated a strong association between this interaction and the error rate for Data Pattern I. For Data Pattern II, $\omega^2_{partial}$ value (.014) signified a weak association. The interaction is presented graphically in Figure G2-6. The means and standard deviations are summarized in Table G2-7.

As shown in Figure G2-6, the performances of the four methods depended on the degree of separation between the two populations' means. The further the two populations separated, the lower was G2 error rate regardless which method was used. The performances of LPM, LDF, and LR were similar for either data pattern. However, KM performed differently for the two data patterns, also differently from LPM, LDF, and LR.

In Table G2-7, the mean G2 error rates for LPM were slightly lower than those for LDF and LR regardless the degree of group separation. LDF and LR performed similarly for three group separation levels. For both data patterns, KM performed poorly compared to the other three methods. The impact of group separation on the performance of KM was larger for Data Pattern I than for Data Pattern II. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: LPM was a viable alternative regardless of the degree of group separation whereas KM was not the method of choice. Group separation had little impact on the relative efficiency of LR over LDF.

Method by sample representativeness interaction. This two-way interaction effect accounted for 1.27% ($\eta^2 = .0127$) and 0.49% ($\eta^2 = .0049$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.264) for Data Pattern I indicated a strong association between this interaction and the error rate. However, for Data Pattern II, $\omega^2_{partial}$ value of .113 signified a moderate association. The interaction is presented graphically in Figure G2-7. The means and standard deviations are summarized in Table G2-8.

As shown in Figure G2-7, the performances of the four methods depended on sample representativeness. The performances of LPM, LDF, and LR were similar for each data pattern but different from KM. The mean G2 error rates were the highest when Group 1 was 20% over-sampled; the mean error rates were the lowest when Group 1 was 20% under-sampled. Although the performance of KM was different for the two data patterns, the mean error rates of KM under the three conditions of sample representativeness were similar.

In Table G2-8, for both data patterns, the mean G2 error rates of KM were higher than those of the other three methods regardless of sample representativeness. KM was not as good as the other three methods. However, unlike the other three methods, KM performed similarly at all levels of sample representativeness. The mean error rates were lower under the condition of 20% Group 1 over-sampled than those of the equal condition. Meanwhile, the mean error rates of the 20% under-sampled condition were slightly higher than those of the equal condition. The discrepancies between η^2 and $\omega^2_{partial}$ values for the two data patterns were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: LPM was the best method. The impact of sample representativeness on predictive performances was similar for LPM, LDF, and LR, but different from KM. Although KM did not perform as well as the other three methods, its performance was least influenced by sample representativeness.

Three-way Method by Data Property Interactions

Five three-way interactions regarding the method factor are discussed in this section.

They were (a) method by population proportion by equality of covariance matrices interaction, (b) method by population proportion by group separation interaction, (c) method by population proportion by sample representativeness interaction, (d) method by equality of covariance matrices by group separation interaction, and (e) method by group separation by sample representativeness interaction. For Data Pattern I, interactions (a), (b), (c) and (e) from the list above were considered practically significant based on the criterion of $\omega^2_{partial} > .06$. However, for Data Pattern II, only three interactions [i.e., (b), (c) and (d) from the list above] were considered practically significant.

Method by population proportion by equality of covariance matrices interaction. This three-way interaction effect accounted for 0.34% ($\eta^2 = .0034$) and 0.08% ($\eta^2 = .0008$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ value (.087) for Data Pattern I indicated a moderate association between this interaction and the error rate. However, for Data Pattern II, $\omega^2_{partial}$ (.019) signified a weak association. The interaction is presented graphically in Figure G2-8. The means and standard deviations are summarized in Table G2-9.

As shown in Figure G2-8, the interaction profiles for LPM, LDF, and LR were similar for the two data Patterns. The mean G2 error rates decreased as the population proportions deviated from the 0.5:0.5 split, regardless of the equality of covariance matrices. For all levels of population proportions, the mean error rate was the highest (i.e., a negative effect) if Population #1 had smaller covariances and the lowest (i.e., a positive effect) if Population #1 had larger covariances. The impact of inequality of covariance matrices on G2 error rate was small when the population proportions were 0.1:0.9. The impact increased as the population proportions approached the 0.5:0.5 split.

Figure G2-8(d) illustrates the population proportion by equality of covariance matrices interaction under KM. The interaction profiles were different for the two data patterns. In Table G2-9, the mean G2 error rates of KM increased as the population proportions deviated from the 0.5:0.5 split, regardless of equality of covariance matrices. For all levels of population proportions, the mean error rate was the highest if Population #1 had smaller covariances and the lowest if Population #1 had larger covariances. The impact of inequality of covariance matrices on G2 error rate was small when the population proportions were 0.1:0.9. The impact increased as the population proportions approached the 0.5:0.5 split. The mean error rates of KM for Data Pattern I were lower than those for Data Pattern II. Yet the impact of equality of covariance matrices on error rates was greater in Data Pattern I than in Data Pattern II. The mean G2 error rates of KM were higher than those of LPM, LDF, or LR in most conditions for Data Pattern I. When the population proportions were 0.5:0.5 and either both covariances were equal or Population #1 had larger covariances, KM performed better than the other three methods. For Data Pattern II, KM performed poorly compared to the other three methods in all combinations of population proportion by equality of covariance matrices. KM performed as well as the other three methods when the population proportions were 0.5:0.5 and covariances were either equal or Population #1 had larger covariances.

Method by population proportion by group separation interaction. This three-way interaction effect accounted for 2.26% ($\eta^2 = .0226$) and 0.41% ($\eta^2 = .0041$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.389) for Data Pattern I indicated a strong association between this interaction and the error rate. However, for Data Pattern II, the $\omega^2_{partial}$ value (.097) signified a moderate association. The interaction is presented graphically in Figure G2-9. Separate plots of mean G2 error rates due to the population proportion by group separation interaction are presented in (a), (b), (c), and (d) for the four methods. The means and standard deviations are summarized in Table G2-10.

As shown in Figure G2-9, performances of the four methods in predicting the membership of Population #2 depended on the joint conditions of population proportions and group separation. The interaction profiles for LPM, LDF, and LR were similar for the two data Patterns. Yet they were different from KM's.

In Figures G2-9(a), (b), and (c), the mean error rates were the highest when the population proportions were 0.5:0.5 and the lowest under the 0.1:0.9 condition, regardless of degrees of group separation. The impact of group separation on G2 error rate was inconsistent under different population proportions. The inconsistency was not only in magnitude, but also in direction. When population proportions were 0.1:0.9, the mean G2 error rates were extremely low regardless of degrees of group separation. The mean error rate decreased as the group separation decreased. When population proportions equaled 0.25:0.75, the mean error rates were slightly higher than when they were 0.1:0.9. The mean error rate increased slightly as the group separation decreased from 6.7 to 2.2; it decreased only slightly when the group separation decreased from 2.2 to 0.7. The mean error rate increased as the group separation decreased, when the population proportions were 0.5:0.5. The performances of LPM, LDF and LR were comparable when the population proportions were 0.5:0.5.

Figure G2-9(d) presents the population proportion by group separation interaction under KM. KM performed as well as the other three methods when the population proportions were 0.5:0.5, especially when the two populations were not well separated (Table G2-10). For Data Pattern II, the mean error rates of KM were substantially higher than those of the other three methods. Different conditions of population proportions had little impact on the performance of

KM when group separation (d^2) was 2.2 or 0.7. When $d^2 = 6.7$, the mean G2 error rate increased as the population proportions deviated from 0.5:0.5.

Results indicated that LPM was the best method especially when the population proportions deviated from 0.5:0.5. The impact of group separation on performances of LPM, LDF, and LR depended on population proportions. Group separation had little impact on the relative efficiency of LR over LDF. KM performed as well as the other three methods when the two populations were not well separated and the population proportions were 0.5:0.5.

Method by population proportion by sample representativeness interaction. This three-way interaction effect accounted for 0.70% ($\eta^2 = .007$) and 0.41% ($\eta^2 = .0041$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.165) for Data Pattern I indicated a strong association between this interaction and the error rate. For Data Pattern II, however, the $\omega^2_{partial}$ value (.097) signified a moderate association. The interaction is presented graphically in Figure G2-10. The means and standard deviations are summarized in Table G2-11.

As shown in Figure G2-10, the performances of the four methods depended on the combination of population proportions and sample representativeness. The interaction profiles for LPM, LDF, and LR were similar for the two data patterns. Yet they were different from those of KM.

In Figures G2-10(a)-(c), the mean G2 error rates decreased as the population proportions deviated from 0.5:0.5 regardless of sample representativeness. The mean G2 error rates were the lowest when Group 1 was 20% under-sampled while the mean error rates were the highest when Group 1 was 20% over-sampled regardless of population proportions. The performances of LPM, LDF and LR were comparable when the population proportions were 0.5:0.5. The performances of LDF and LR were similar in all combinations of population proportions and sample representativeness.

Figure G2-10(d) presents the population proportion by sample representativeness interaction under KM. The interaction profiles were different for the two data patterns. In Table G2-11, for Data Pattern I, the mean G2 error rates increased as the population proportions deviated from 0.5:0.5 regardless of sample representativeness. Unlike the results of the other three methods, the mean G2 error rates of KM were the highest when Group 1 was 20% under-sampled and the lowest when Group 1 was 20% over-sampled, regardless of population proportions. KM performed as well as the other three methods only when the population proportions were 0.5:0.5 and Group 1 was either 20% over-sampled or "equally" sampled. For Data Pattern II, the mean G2 error rates increased as the population proportions deviated from 0.5:0.5 regardless of sample representativeness. However, sample representativeness had little impact on KM's performance.

Results indicated that, for LPM, LDF, and LR, the impact of sample representativeness on G2 error rates depended on population proportions. For KM, the impact was consistent for each level of population proportions. LPM performed the best when the population proportions deviated from 0.5:0.5. The performances of LDF and LR were similar. Depending on the data pattern, KM could be a viable alternative when population proportions were 0.5:0.50 and sample representativeness was either "equal" or Group 1 was over-sampled.

Method by equality of covariance matrices by group separation interaction. This three-way interaction effect accounted for 0.06% ($\eta^2 = .0006$) and 0.27% ($\eta^2 = .0027$) of the sample variance of G2 error rates for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.016) for Data Pattern I indicated a weak association between this interaction and the error rate. However, for Data Pattern II, the $\omega^2_{partial}$ value (.066) signified a moderate association. The interaction is presented graphically in Figure G2-11. The means and standard deviations are summarized in Table G2-12.

As shown in Figure G2-11, the performances of the four methods depended on the combination of equality of covariance matrices and group separation. The interaction profiles for LPM, LDF, and LR were similar for the two data patterns. Yet they were different from those of KM.

In Figures G2-11(a)-(c), these interaction profiles for the three methods were similar. The mean G2 error rates increased as the separation between the two populations decreased regardless of the equality of covariance matrices. For each level of group separation, the performances of the three methods were the best, when Population #2's covariances were smaller than those of Population #1.

Figure G2-11(d) presents the equality of covariance matrices by group separation interaction under KM. The interaction profiles were different for the two data patterns. The mean G2 error rates of KM increased as the separation between the two populations decreased regardless of the equality of covariance matrices (also Table G2-12). At each level of group separation, the performance of KM was the best when the covariances of Population #2 were smaller than those of Population #1. The mean G2 error rates of KM for Data Pattern I were lower than those for Data Pattern II. Impacts of this interaction on G2 error rates of KM's were stronger for Data Pattern I than for Data Pattern II. KM performed poorly compared to the other three methods in all combinations of equality of covariance matrices and group separation.

Results indicated that the further the two populations separated, the lower was the G2 error rate, regardless which statistical method was used. The mean error rate was lower if Population #2 had smaller covariances than the other two conditions, regardless of the degree of group separation. LPM was the best method in all combinations of equality of covariance matrices and group separation. KM's performance was the worst. LDF and LR performed differently under unequal covariance matrices. LR performed better than LDF when Population #2 had larger covariances. However, LDF performed better than LR when Population #2 had smaller covariances.

Method by group separation by sample representativeness interaction. This three-way interaction effect accounted for 0.25% ($\eta^2 = .0025$) and 0.13% ($\eta^2 = .0013$) of the sample variance of G2 error rate for Data Patterns I and II, respectively. $\omega^2_{partial}$ (.067) for Data Pattern I indicated a moderate association between this interaction and the error rate. However, for Data Pattern II, the $\omega^2_{partial}$ value (.032) signified a weak association. The interaction is presented graphically in Figure G2-12. The means and standard deviations are summarized in Table G2-13.

As shown in Figure G2-12, the performances of the four methods depended on the combination of group separation and sample representativeness. The interaction profiles for LPM, LDF, and LR were similar for the two data patterns. Yet they were different from those of KM.

In Figures G2-12(a)-(c), the interaction profiles for LPM, LDF, and LR were similar. The mean G2 error rates increased as the group separation decreased regardless of sample representativeness. For each level of group separation, the mean error rate was the highest when Group 1 was 20% over-sampled and the lowest when Group 1 was 20% under-sampled. The impact of sample representativeness on the error rate increased as the group separation decreased. The differences between the equal sample represenativeness and the other two conditions increased as the degree of group separation decreased.

Figure G2-12(d) presents the group separation by sample representativeness interaction under KM. The interaction profiles were different for the two data patterns. For both data patterns, the mean G2 error rates of KM increased as the group separation decreased regardless of sample representativeness. Unlike the other three methods, the mean error rate of KM was the lowest when Group 1 was 20% over-sampled and highest when Group 1 was 20% under-sampled, regardless of group separation. The impact of sample representativeness on the error rates of KM was consistent across the three levels of group separation. The mean G2 error rates of KM for Data Pattern II were higher than those for Data Pattern I. Yet the impact of group separation on G2 error rates was larger for Data Pattern I than for Data Pattern II. For both data patterns, KM performed poorly, compared to the other three methods.

Results indicated that the group separation by sample representativeness interaction for LPM, LDF, and LR were similar. Yet they were different from those of KM. For LPM, LDF, and LR, the impact of sample representativeness on G2 error rate increased as the separation between the two populations decreased. For KM, the impact was consistent for each level of the group separation. LPM yielded the lowest mean G2 error rates in all combinations of group separation and sample representativeness. The performances of LDF and LR were comparable. Although KM performed poorly compared to the other three methods, KM was least affected by sample representativeness.

7. Results of Total Error Rates

Total error rate was the proportion of observations in the test sample that were misclassified. The means and standard deviations of the error rates of 200 replications for each combination of levels of the five factors for Data Pattern I are summarized in Table T-1. Similarly, the means and standard deviations for Data Pattern II are presented in Table T-2. The results of ANOVA on Total error rates for Data Pattern I are summarized in Table T-3. Out of 31 main and interaction effects, 23 effects were statistically significant at the .0016 level. The ANOVA results for Total error rate for Data Pattern II are presented in Table T-4. Twenty-seven effects were statistically significant at the .0016 level for Data Pattern II.

Data Property Main Effects

The effects of the five "between-subjects" main factors on Total error rate were statistically significant at the .0016 level for both data patterns. However, only three of the five factors were found to be practically significant using the criterion of $\omega^2_{partial} > .06$ for Data Pattern I (see Table T-3). These three factors were Factor 1 (population proportion), Factor 2 (equality of covariance matrices), and Factor 3 (group separation). For Data Pattern II, only two factors (i.e., Factor 1 and Factor 3) had a $\omega^2_{partial}$ value greater than .06 (see Table T-4).

Population Proportion

For Data Pattern I, Factor 1 (population proportion) explained 6.68% ($\eta^2 = .0668$) of the sample variance of Total error rates. This factor signified a strong association ($\omega^2_{partial} = .277$) with Total error rates. The mean error rates for the three levels of population proportions were .146, .182, and .216. Results from the Newman-Keuls procedure indicated that the three mean error rates were statistically significantly different from each other. The mean Total error rate (.146) was the lowest under 0.1:0.9 population proportions and the highest (.216) under the 0.5:0.5 condition.

Similarly, for Data Pattern II, this factor explained 5.37% ($\eta^2 = .0537$) of the sample variance of Total error rates. The $\omega^2_{partial}$ value (.385) signified a strong association between population proportions and Total error rates. Results from the Newman-Keuls procedure indicated that the mean error rate (.275) for the 0.5:0.5 population proportion condition was statistically significantly higher than that (.241) for the 0.25:0.75 condition. The mean error rate (.185) was the lowest for the 0.1:0.9 condition; it was statistically significantly lower than the means for the other two conditions. Based on the results from both data patterns, Total error rate decreased as the proportions of Population #1 and Population #2 deviated from the 0.5:0.5 split.

Equality of Covariance Matrices

Factor 2 (equality of covariance matrices) had three levels: (a) equal covariance matrices, (b) Population #1 had smaller covariances that were one-fourth of the covariances of Population #2, and (c) Population #1 had larger covariances that were four times of the covariances of Population #2. For Data Pattern I, this factor explained 1.23% ($\eta^2 = .0123$) of the sample variance of Total error rates. The factor exhibited a moderate association ($\omega^2_{partial}$ = .0657) with the Total error rate. Results from the Newman-Keuls procedure indicated that the mean Total error rates for the three levels were statistically significantly different from each other. When Population #1 had larger covariances, the mean error rate was the lowest (.164). The mean error rate was .173 under the equal covariance matrices condition. The mean error rate was the highest (.191) when Population #1 had smaller covariances.

For Data Pattern II, the equality of covariance matrices factor explained 0.36% ($\eta^2 = .0036$) of sample variance of Total error rates. This factor had a weak association ($\omega^2_{partial}$ = .040) with Total error rates. Results from the Newman-Keuls procedure indicated that the mean Total error rates for the three levels were statistically significantly different from each other. When Population #1 had larger covariances, the mean error rate (.220) was the lowest. The mean error rate (.242) was the highest for the equal covariance matrices condition. When Population #1 had smaller covariances, the mean error rate was .238.

This factor was statistically significant at $\alpha = .0016$ for both data patterns; the absence of strong associations implied that the differences in mean Total error rates for the three levels of equality of covariance matrices were practically insignificant. In other words, a ratio of 1:4 (or

4:1) for heterogeneity of covariance matrices exhibited only a small impact on Total error rates.

Group Separation

Factor 3 (group separation) explained 37.97% ($\eta^2 = .3797$) of the sample variance of Total error rates for Data Pattern I. A strong association ($\omega^2_{partial} = .685$) between this factor and Total error rates was detected. Results from the Newman-Keuls procedure indicated that three mean error rates of this factor differed from each other. When the Mahalanobis distance (d^2) between the two populations' means was 6.709, the mean error rate was .093 (the lowest). As d^2 decreased, the mean error rate increased. Under the condition of $d^2 = 2.236$, the mean error rate was .192. The mean error rate was .259 (the highest) when d^2 was 0.745.

For Data Pattern II, similar results were obtained. This factor accounted for 11.19% ($\eta^2 = .1119$) of the sample variance of the Total error rate. The association between this factor and the error rate was strong ($\omega^2_{partial} = .567$). Similar to the results obtained from Data Pattern I, as d^2 decreased, the mean Total error rate increased. The mean error rate increased from .163 (when $d^2 = 6.785$) to .245 (when $d^2 = 2.262$) and further increased to .293 (when $d^2 = 0.754$).

Results from both data patterns led to the conclusion that the further the two populations separated, the lower was Total error rate. These findings were expected because the less overlapping of the populations, the less likely that sample observations would be misclassified.

Sample representativeness

Factor 4 (sample representativeness) had three levels: (a) Group 1 (i.e., sample comprised of observations from Population #1) was 20% over-sampled, (b) sample proportions equal to the population proportions, and (c) Group 1 was 20% under-sampled. This factor was statistically significant at $\alpha = .0016$ for both data patterns. However, for Data Pattern I, only 0.06% ($\eta^2 = .0006$) of the sample variance of Total error rates was explained by this factor. A negligible association ($\omega^2_{partial}$ = .004) was found between this factor and the Total error rate. Results from the Newman-Keuls procedure indicated that the mean error rate (.180) of the "20% over-sampled" condition was statistically significantly smaller than that (.185) obtained under the "20% under-sampled" condition. These two mean error rates were statistically significantly higher than that (.179) of the "equal" condition.

Similarly, for Data Pattern II, this factor explained 0.02% (η^2 = .0002) of the sample variance of Total error rates. $\omega^2_{partial}$ (.002) indicated a negligible association between group representativeness and Total error rates. Results from the Newman-Keuls procedure showed that the mean error rate (.233) of the "20% over-sampled" condition was statistically significantly smaller than that (.236) of the "20% under-sampled" condition. Both mean error rates were statistically significantly higher than that (.231) of the "equal" condition.

Although this factor was statistically significant at α = .0016 for both data patterns, an absence of strong associations implied that the differences in mean Total error rates were not likely to be attributable to the three levels of sample representativeness.

Sample size

Two levels of sample size (i.e., 200 and 400) were considered in this study. For Data Pattern I, less than 0.01% ($\eta^2 < .0001$) of the sample variance of Total error rate was explained by this factor. This factor was virtually not associated with the error rate ($\omega^2_{partial} < .0001$). The mean error rates for sample sizes of 200 and 400 were .182 and .181, respectively. Similar results were obtained from Data Pattern II. This factor accounted for 0.01% (η^2 = .0001) of the sample variance of Total error rate. The $\omega^2_{partial}$ value (.0008) indicated a virtually non-existent association between this factor and Total error rate. The mean error rates for sample sizes of 200 and 400 were .234 and .233, respectively.

Although this factor was statistically significant at $\alpha = .0016$ for both data patterns, evidence of a non-existent association implied that the differences in mean error rates were not likely to be attributable to the two levels of sample size. By examining Tables T-1 and T-2, one notices that the standard deviations of Total error rates for sample size of 400 were smaller than those for sample size of 200. In other words, with a larger sample size, one obtains a more efficient estimate of Total error rates.

Two-way Interaction among Data Property Factors

Judged against the $\omega^2_{partial} > .06$ criterion, the sole two-way interaction considered practically significant was the population proportion by group separation interaction. This interaction is the only result discussed here.

Population Proportion by Group Separation Interaction

This interaction was statistically significant at $\alpha = .0016$ for both data patterns. For Data Pattern I, it accounted for 5.31% ($\eta^2 = .0531$) of the sample variance of the Total error rate and had a large association ($\omega^2_{partial} = .233$) with the error rate. For Data Pattern II, this factor explained 3.57% ($\eta^2 = .0357$) of the sample variance of the error rate and also exhibited a large association ($\omega^2_{partial} = .294$). This interaction is graphically presented in Figure T-1.

As shown in Figure T-1, the interaction profiles for the two data patterns were similar. The Total error rates were slightly lower for Data Pattern I than for Data Pattern II. When the separation of the two populations was large (i.e., $d^2 = 6.7$), the population proportion factor had a smaller impact on the Total error rate. However, when the separation was smaller (i.e., $d^2 = 2.2$ or 0.7), the impact of population proportion on the error rate increased. The differences in the mean Total error rates between 0.50:05 and 0.1:0.9 population proportion conditions increased from .001 (when $d^2 = 6.7$) to .18 (when $d^2 = 0.7$).

Results from both data patterns indicated that the impact of population proportions on Total error rates was the strongest when the two populations were not well separated (i.e., $d^2 = 0.7$). When the two populations were well separated (i.e., $d^2 = 6.7$), differences in population proportions had little impact on Total error rates.

Effects Concerning Statistical Methods

For Data Pattern I, four effects concerning the method (i.e., the "within-subjects") factor were considered practically significant according to the criterion of $\omega^2_{partial} > .06$. These five effects were: (a) the main effect of methods, (b) method by population proportion interaction, (c) method by equality of covariance matrices interaction, (d) method by group separation interaction, and (e) method by population proportion by group separation interaction. For Data Pattern II, four of the five effects listed above were considered practically significant except for the method by equality of covariance matrices interaction.

Main Effect of the Method Factor

The four statistical methods (i.e., LPM, LDF, LR, and KM) were treated as levels of the "method" factor. This factor was statistically significant at $\alpha = .0016$ and accounted for 18.49% ($\eta^2 = .1849$) and 68.48% ($\eta^2 = .6848$) of the sample variance of Total error rates for Data Patterns I and II, respectively. $\omega^2_{partial}$ equaled .829 and .947 for Data Patterns I and II, respectively indicating a strong association.

The mean Total error rate for LPM was .157 for Data Pattern I, and .158 for Data Pattern II. The mean error rates for LDF were .152 and .155 for Data Patterns I and II, respectively. The mean error rate for LR was .153 for Data Pattern I and .157 for Data Pattern II. The mean error rates for KM were .264 and .464 for Data Patterns I and II, respectively. The Dunn-Šidák procedure was applied to further examine pairwise mean differences among the four methods. Results from the Dunn-Šidák procedure indicated that the means Total error rates were statistically significantly different from each other for both data patterns.

In general, KM yielded the highest mean Total error rate and LDF the lowest. The mean error rates for LPM were only slightly higher than those for LR, even though the differences were statistically significant.

Two-way Method by Data Property Interaction

For both data patterns, the method factor was found statistically significantly interacting with all five data property factors at $\alpha = .0016$. For Data Pattern I, only 3 two-way interactions were considered practically significant according to the criterion of $\omega^2_{partial} > .06$. These three interactions were (a) method by population proportions interaction, (b) method by equality of covariance matrices interaction, and (c) method by group separation interaction. For Data Pattern II, only the method by population proportions interaction and the method by group separation interaction were considered practically significant. These three interactions are discussed in greater details below.

Method by population proportion interaction. This two-way interaction accounted for 18.27% ($\eta^2 = .1827$) and 4.34% ($\eta^2 = .0434$) of the sample variance of Total error rates for Data Patterns I and II, respectively. $\omega^2_{partial}$ values (.827 for Data Pattern I and .533 for Data Pattern II) indicated a strong association between this interaction and the error rate. The interaction is graphically presented in Figure T-2. Similar interaction profiles were found for both data patterns. The means and standard deviations for the interaction are summarized in Table T-5.

As shown in Figure T-2, the interaction profiles for the two data patterns were similar to the profiles shown in Figure G2-4. The performances of the four methods depended on sample representativeness. For each data patter, the performances of LPM, LDF, and LR were similar but different from KM. The mean Total error rates of these three methods were the lowest when the population proportions were 0.1:0.9 and the highest under the 0.5:0.5 split.

Although the performance of KM was different for the two data patterns, the mean error rates were the lowest under the 0.5:0.5 condition and the highest under the 0.1:0.9 condition. According to Table T-5, for Data Pattern I, KM slightly outperformed the other three methods only when the population proportions were 0.5:0.5. When population proportions were 0.25:0.75 and 0.1:0.9, Total error rates of KM were substantially higher than those for the other three methods. For Data Pattern II, KM performed poorly in all three population proportion conditions, compared to the other three methods. The discrepancies between η^2 and $\omega^2_{partial}$ values were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: when the population proportions were extreme (i.e., 0.1:0.9 or even 0.25:0.75), LDF was the best method in predicting group memberships. KM was not a method of choice when population proportions were extreme. When population proportions were 0.5:0.5, the performances of LPM, LDF, and LR were almost identical; KM performed as well as the other three methods only for Data Pattern I.

Method by equality of covariance matrices interaction. This two-way interaction accounted for 0.46% ($\eta^2 = .0046$) and 0.04% ($\eta^2 = .0004$) of the sample variance of Total error rates for Data Patterns I and II, respectively. The $\omega^2_{partial}$ value (.107 for Data Pattern I) indicated a moderate association between this interaction and the error rate. However, for Data Pattern II, the $\omega^2_{partial}$ value (.011) signified a weak association. The interaction is graphically presented in Figure T-3. A similar interaction profile was found for both data patterns. The means and standard deviations for the interaction are summarized in Table T-6.

As shown in Figure T-3, unequal covariance matrices conditions had a small impact on the differential performances of LPM, LDF, and LR. Yet, the mean Total error rates of KM depended on the condition of equality of covariance matrices. The performances of LPM, LDF, and LR were similar for each data pattern. However, KM performed differently for the two data patterns and also from LPM, LDF, and LR.

All four methods performed slightly better when the covariances of Population #1 were four times of those of Population #2. KM performed poorly compared to the other three methods regardless of the conditions of equality of covariance matrices (see Table T-6). Total error rates of KM for Data Pattern I were lower than those of Data Pattern II. The discrepancies between η^2 and $\omega^2_{partial}$ values were caused mainly by the inconsistent performance of KM for the two data patterns.

Results led to the following conclusions: all four methods performed better when Population #1's covariances were larger than Population #2's. LDF was the best method regardless of the condition of equality of covariance matrices whereas KM was the poorest.

Method by group separation interaction. This two-way interaction accounted for 0.93% ($\eta^2 = .0093$) of the Total error rate variance for Data Pattern I, and 0.79% ($\eta^2 = .0079$) for Data Pattern II. A strong association between this interaction and the error rate was detected ($\omega^2_{partial} = .195$ and .173 for Data Patterns I and II, respectively). The interaction is graphically presented in Figure T-4 and the means and standard deviations of Total error rate are summarized in Table T-7.

As shown in Figure T-4, the performances of the four methods depended on the degree of separation between the two populations' means. The further the two populations separated, the lower was the Total error rate regardless which method was used. The performances of LPM, LDF, and LR were similar for each data pattern. However, KM performed differently for the two data patterns, also from LPM, LDF, and LR.

In Table T-7, for both data patterns, KM performed poorly compared to the other three methods regardless of the degree of group separation. The mean Total error rates of KM were lower for Data Pattern I than for Data Pattern II. The impact of group separation on Total error rates was larger for Data Pattern I than for Data Pattern II. The differences among the mean Total error rates of the three levels of group separation were larger for Data Pattern I than for Data Pattern II.

Results led to the following conclusions: LPM performed as well as LDF or LR when the two populations were not well separated. KM did not perform as well as the other three methods.

Three-way Method by Data Property Interaction

The method by population proportion by group separation interaction was the only three-way interaction considered practically significant according to the criterion of $\omega^2_{partial} > .06$. In addition to this three-way interaction, the method by population proportion by equality of covariance matrices interaction is also included to contrast the results of this study with the findings from Fan and Wang (1999). They are the only results discussed here.

Method by population proportion by group separation interaction. This interaction explained 1.85% ($\eta^2 = .0185$) and 0.26% ($\eta^2 = .0026$) of the sample variance of Total error rates for Data Patterns I and II, respectively. A strong association between this interaction and the error rate was detected ($\omega^2_{partial} = .326$) for Data Pattern I. However, for Data Pattern II, only a moderate association ($\omega^2_{partial} = .065$) was found. The means and standard deviations of the error rate for the three-way interaction are summarized in Table T-8. The interaction is graphically presented in Figure T-5.

As shown in Figure T-5, the performances of the four methods in predicting the group membership depended on the combinations of population proportions and group separation. The interaction profiles for LPM, LDF, and LR were similar for the two data Patterns. Yet they were different from those of KM.

As shown in Figures T-5(a)-(c), the mean Total error rates increased as the group separation decreased regardless of population proportions. The mean Total error rates were the highest when population proportions were 0.5:0.5 and the lowest when proportions were 0.1:0.9, regardless of the degree of group separation. However, the differences among the mean error rates, at the three levels of population proportions, became larger as the separation between the two populations decreased.

Figure T-5(d) illustrates the population proportion by group separation interaction under KM. The interaction profiles were different for the two data patterns. For Data Pattern I, the mean Total error rates increased as the separation between the two populations decreased, regardless of population proportions. Meanwhile, the mean error rates increased as the population proportions deviated from the 0.5:0.5 split, regardless of group separation. KM performed as well as the other three methods when the population proportions were 0.5:0.5 and, especially, when the two populations were not well separated (see Table T-8). For Data Pattern II, the mean error rates of KM were substantially higher than those of the other three methods. Different population proportions had little impact on KM's performances, when group separations (d^2) were 2.2 and 0.7. When $d^2 = 6.7$, the mean Total error rates increased as the population proportions deviated from 0.5:0.5.

Results indicated that LPM performed as well as LDF and LR when either the population proportions were within the range (i.e., 0.25:0.75 – 0.5:0.5) or the two populations were not well separated (i.e., $d^2 = 2.2$ or 0.7). LR did not outperform LDF when the population proportions were 0.1:0.9. KM performed as well as the other three methods when the population proportions were 0.5:0.5 and the two populations were not well separated.

Method by population proportion by equality of covariance matrices interaction. This three-way interaction explained 0.23% ($\eta^2 = .0023$) and 0.01% ($\eta^2 = .0001$) of the sample variance of Total error rates for Data Patterns I and II, respectively. This interaction exhibited a weak association ($\omega^2_{partial} = .058$ and .003 for Data Patterns I and II) with the Total error rate. This interaction is included here in order to contrast the results of this study with the findings from Fan and Wang (1999). The means and standard deviations of Total error rate for the three-way interaction are summarized in Table T-9.

In Table T-9, Total error rates of LPM, LDF, and LR were comparable. The performance of LPM was not as good as the other two methods when population proportions deviated from 0.5:0.5. The performance of KM was different for the two data patterns. For Data Pattern I, KM performed slightly better than the other three methods when population proportions were 0.5:0.5. For Data Pattern II, KM did not perform as well as the other three methods in all combinations of population proportion and equality of covariance matrices. When population proportions were either 0.1:0.9 or 0.25:0.75, smaller covariances in Population #1 had a negative impact on the Total error rates whereas larger covariance in Population #1 had a positive impact. However, when population proportions were 0.5:0.5, both conditions of covariance matrices had a positive impact on the error rates.

Results indicated that (a) when population proportions were 0.5:0.5, violation of homogeneity of covariance matrices did improve the accuracy in predicting group membership; (b) if the Total error rate was the main concern, any of LPM, LDF, or LR methods performed as well as the others; and (c) KM was a viable alternative when population proportions were 0.5:0.5 and the Total error rate was the objective of the research.

8. Implications for Educational Researchers

Classification enables men and women to make sense of the information they encounter. Aldenderfer and Blashfield (1984) state,

> [c]lassification is also a fundamental process of the practice of science since classification systems contain the concepts necessary for the development of theories within a science. (p. 7)

Classification is closely related to another important human activity—prediction. In this study, the accuracies of predicting two-group membership by K-means clustering were compared with those derived from linear probability modeling, linear discriminant function, and logistic regression under various data properties. Three predictive error rates (Group 1, Group 2, and Total) provided the basis for comparisons.

Findings in this study echoed Gilbert's (1969) conclusion that moderate violation of homogeneity of covariance matrices assumption had only a mild impact on error rates. The use of a ratio of 4 in the two conditions of heterogeneity of covariance matrices exhibited only a small impact on the three error rates. The direction of impact of heterogeneity of covariance matrices on the performances of LPM, LDF, LR and KM depended on population proportions and the type of error rates.

In the field of education, the targeted population proportions are mostly extreme. The accuracy of predicting membership of the smaller population (i.e., Population #1 in this study) is frequently the main focus. The selection of LPM, LDF, or LR in this situation depended on the conditions of heterogeneity of covariance matrices. When population proportions were 0.1:0.9, the condition in which Population #1 had smaller covariances exhibited a negative impact while the condition in which Population #1 had larger covariances had a positive impact on the error rates of LPM and LDF. However, both conditions of heterogeneity of covariance matrices had a positive impact on the performance of LR. LPM was not the method of choice when the accuracy of predicting membership in the smaller population (i.e., Population #1) was the main objective. LR should be selected when Population #1 had smaller covariances; LDF should be selected when Population #1 had larger covariances.

For KM, when unequal covariance matrices of the two populations existed, observations from the population with smaller covariances were less likely to be misclassified than members of the population with larger covariances. When population proportions deviated from 0.5:0.5, the condition in which Population #1 had larger covariances minimized the Group 2 error rate. Consequently, this condition had the lowest Total error rate. When population proportions were 0.5:0.5, both conditions of heterogeneity of covariance matrices minimized the Total error rate, compared to the Total error rate obtained when the two populations had equal covariance matrices. Depending on the data pattern, KM could be an alternative method especially when the population proportions were extreme.

Without the knowledge of population proportions, samples representative of population proportions are hard to obtain in some situations. Kao and McCabe (1991) suggested that equal sample sizes for the two populations should be used in these situations. Findings of this study elaborated on the appropriateness of this suggestion. If the proportions of the two populations were similar, maintaining equal sample sizes would have little impact on individual group's as well as the total error rates. However, if the population proportions were extreme, equal sample sizes implied that Population #1 was over-sampled while Population #2 was under-sampled.

Consequently, for LPM, LDF, and LR, there would be a reduction in Group 1 error rate while an increment in Group 2 error rate, compared to the error rates obtained from a representative sample. Sampling equal number of observations from the underlying populations would favor the prediction of membership for the smaller population, but, had little impact on the Total error rates.

For KM, when Population #1 was over-sampled, there would be an increment in Group 1 error rate while a reduction in Group 2 error rate, compared to the error rate obtained from a representative sample. In other words, sampling equal number of observations from the underlying populations would favor the prediction of membership for the larger population. Consequently, higher Group 1 and lower Total error rates would be obtained in this case, compared to the error rates obtained from a representative sample.

Findings in this study partially supported the proposition that the further the two populations separated, the lower was the error rate. For KM, this proposition correctly described the impact of group separation on Total as well as separate group error rates. For LPM, LDF, and LR, this proposition describes the impact of group separation on only Group 1 and Total error rates. However, the impact of group separation on Group 2 error rate was not fully consistent with the above proposition. When population proportions were 0.5:0.5, the further the two populations separated, the lower was Group 2 error rate. However, when population proportions were 0.25:0.75, the mean error rate was the lowest when $d^2 = 6.7$ and the highest when $d^2 = 2.2$. The irregular impact of group separation on Group 2 error rate was unexpected. Additional studies are needed in order to fully investigate this phenomenon. Findings of this study did not support the notion that increasing group separation should have a negative impact on the superiority of LR over LDF. Increasing group separation may lower the efficiency of LR relative to LDF in parameter estimations (Efron, 1975). However, the inefficiency in parameter estimations did not seem to have a strong impact on the accuracies of prediction within the range of Mahalanobis distance from 0.7 to 6.7.

The present study found similar performances by LPM, LDF, and LR for the two data patterns. This implies that the data property factors had consistent and uniform impacts on the error rates regardless of the data pattern. However, for KM, the performances were different for the two data patterns. This implies that some factor(s) other than those manipulated and investigated in this study had an influence on the performance of KM in predicting two group memberships. One of the possible factors is the mean structures of the two data patterns. For Data Pattern I, all three means of Population #2 were higher than those of Population #1. However, for Data Pattern II, five out of eight means (i.e., X1, X3, X3, X4, and X8) were higher in Population #1 than Population #2; Population #2 had higher means in the remaining three covariates. The differences in mean structure indicated that the two populations overlapped differently for the two data patterns. The pattern of how the two populations overlapped affected the performances of KM, but had little effect on the performances of the other three methods.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of Ph.D. programs in North America. *American Psychologist, 45*(6), 721-734.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-044). Newbury Park, CA: SAGE.

Balakrishnan, P. V., Cooper, M. C., Jacob, V. S., & Lewis, P. A. (1994). A study of the classification capabilities of neural networks using unsupervised learning: A comparison with K-means clustering. *Psychometrika, 59*(4), 509-525.

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior, 25*, 334-348.

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education, 34*, 569-581.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70*, 892-898.

Elmore, P. B., & Woehlke, P. L. (April, 1996). *Research methods employed in "American Educational Research Journal," "Educational Researcher," and "Review of Educational Research" from 1978 to 1995*. Paper presented at the 1996 annual meeting of the American Educational Research Association, New York, USA. (ERIC Document Reproduction Service No. ED 397122).

Fan, X., and Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education, 67*(3), 265-286.

Fraser, M. W., Jensen, J. M., Kiefer, D., & Popuang, C. (1994). Statistical methods for the analysis of critical events. *Social Work Research, 18*(3), 163-177.

Gilbert, E. S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. *Biometrics, 25*, 505-515.

Grubb, W. N., & Tuma, J. (1991). Who gets student aids? Variation in access to aid. *The Review of Higher Education, 14*, 359-382.

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement, 61*, 909-936.

Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. *Multivariate Behavioral Research, 22*, 307-329.

Kallio, R. E. (1995). Factors influencing the college choice decisions of graduate students. *Research in Higher Education, 36*, 109-124.

Kao, T. C., & McCabe, G. P. (1991). Optimal sample allocation for normal discriminant and logistic regression under stratified sampling. *Journal of the American Statistical Association, 86*, 432-436.

Lei, P. W., & Koehly, L. M. (April, 2000). *Linear discriminant analysis versus logistic regression: A comparison of Classification Errors*. Paper presented at the 2000 annual meeting of the American Educational Research Association, New Orleans, USA.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE.

Meshbane, A., & Morris, J. D. (April, 1996). *Predictive discriminant analysis versus logistic regression in two-group classification problem.* Paper presented at the 1996 annual meeting of the American Educational Research Association, New York, USA. (ERIC Document Reproduction Service No. ED400280).

Mooney, C. Z. (1997). *Monte Carlo simulation* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: SAGE.

Press, J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*, 699-705.

Rice, J. C. (1994). Logistic regression: An introduction. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3, pp. 191-146). Greenwich, CT: JAI Press.

Ryan, T. P. (1997). *Modern regression methods.* New York, NY: Wiley.

So, T. S. H. (2002). *Comparing logistic regression, linear discriminant function, and K-mean clustering in two-group membership prediction: A pilot study.* Unpublished manuscript.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Wilson, R. L., & Hardgrave, B. C. (1995). Predicting graduate student success in an MBA program: regression versus classification. *Educational and Psychological Measurement, 55*, 186-195.

Yarnold, P. R., Hart, L. A., & Soltysik, R. C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement, 54*, 73-85.

Table 1

Data Pattern I

Common correlation matrix (R)	X1	X2	X3
X1	1.00		
X2	.30	1.00	
X3	.50	.40	1.00
Mean structure			
μ_1	5.00	5.00	5.00
μ_2	9.00	9.00	9.00

d^2=6.709

Equal variance condition:

	X1	X2	X3
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{(pooled)}$	4.00	4.00	4.00

Unequal variance conditions [where Population #1 has smaller variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	1.60	1.60	1.60	$\sigma^2_{(1)}$	1.23	1.23	1.23	$\sigma^2_{(1)}$	1.08	1.08	1.08
$\sigma^2_{(2)}$	6.40	6.40	6.40	$\sigma^2_{(2)}$	4.92	4.92	4.92	$\sigma^2_{(2)}$	4.32	4.32	4.32

Unequal variance conditions [where Population #1 has larger variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	6.40	6.40	6.40	$\sigma^2_{(1)}$	9.16	9.16	9.16	$\sigma^2_{(1)}$	12.32	12.32	12.32
$\sigma^2_{(2)}$	1.60	1.60	1.60	$\sigma^2_{(2)}$	2.29	2.29	2.29	$\sigma^2_{(2)}$	3.08	3.08	3.08

d^2=2.236

Equal variance condition:

	X1	X2	X3
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{(pooled)}$	12.00	12.00	12.00

Unequal variance conditions [where Population #1 has smaller variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	4.80	4.80	4.80	$\sigma^2_{(1)}$	3.69	3.69	3.69	$\sigma^2_{(1)}$	3.24	3.24	3.24
$\sigma^2_{(2)}$	19.2	19.2	19.2	$\sigma^2_{(2)}$	14.76	14.76	14.76	$\sigma^2_{(2)}$	12.96	12.96	12.96

Unequal variance conditions [where Population #1 has larger variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	19.2	19.2	19.2	$\sigma^2_{(1)}$	27.44	27.44	27.44	$\sigma^2_{(1)}$	36.92	36.92	36.92
$\sigma^2_{(2)}$	4.80	4.80	4.80	$\sigma^2_{(2)}$	6.86	6.86	6.86	$\sigma^2_{(2)}$	9.23	9.23	9.23

---

d^2=0.745

Equal variance condition:

	X1	X2	X3
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{(pooled)}$	36.00	36.00	36.00

Unequal variance conditions [where Population #1 has smaller variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	14.4	14.4	14.4	$\sigma^2_{(1)}$	11.08	11.08	11.08	$\sigma^2_{(1)}$	9.73	9.73	9.73
$\sigma^2_{(2)}$	57.6	57.6	57.6	$\sigma^2_{(2)}$	44.32	44.32	44.32	$\sigma^2_{(2)}$	38.92	38.92	38.92

Unequal variance conditions [where Population #1 has larger variances]:

Population proportions: .50:.50				Population proportions: .25:.75				Population proportions: .10:.90			
$\sigma^2_{(1)}$	57.6	57.6	57.6	$\sigma^2_{(1)}$	82.28	82.28	82.28	$\sigma^2_{(1)}$	110.76	110.76	110.76
$\sigma^2_{(2)}$	14.4	14.4	14.4	$\sigma^2_{(2)}$	20.57	20.57	20.57	$\sigma^2_{(2)}$	27.69	27.69	27.69

Note. d^2 is group separation measured in Mahalanobis distance: $(\mu_1-\mu_2)'\Sigma^{-1}_{pooled}(\mu_1-\mu_2)$, where $\Sigma_{pooled}=\sigma_{pooled}\times R\times\sigma_{pooled}$, and σ_{pooled} is a diagonal matrix with pooled standard deviations of the variables in the diagonal.

Table 2

Data Pattern II

Common correlation matrix (R)								
	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.00							
X2	.45	1.00						
X3	.05	.25	1.00					
X4	.35	.05	.25	1.00				
X5	.35	.10	.35	.55	1.00			
X6	.05	.25	.50	.15	.40	1.00		
X7	-.35	.05	.40	.15	.30	.41	1.00	
X8	.30	.30	.50	.35	.60	.50	.45	1.00
Mean structure								
μ_1	12.50	15.00	15.95	12.65	12.15	14.15	18.20	15.20
μ_2	11.40	14.25	15.00	11.30	12.90	15.00	19.20	14.50

d^2=6.785

Equal variance condition:

	X1	X2	X3	X4	X5	X6	X7	X8
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{(\text{pooled})}$	1.00	2.00	2.00	1.50	1.20	2.00	2.50	2.00

Unequal variance condition [where Population #1 has smaller variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} / 2.5$	$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} / 3.25$	$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} / 3.7$
$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} \times 4/2.5$	$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} \times 4/3.25$	$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} \times 4/3.7$

Unequal variance condition [where Population #1 has larger variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} \times 4/2.5$	$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} \times 4/1.75$	$\sigma^2_{(1)} = \sigma^2_{(\text{pooled})} \times 4/1.3$
$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} / 2.5$	$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} / 1.75$	$\sigma^2_{(2)} = \sigma^2_{(\text{pooled})} / 1.3$

d^2=2.262

Equal variance condition:

	X1	X2	X3	X4	X5	X6	X7	X8
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{\text{(pooled)}}$	3.00	6.00	6.00	4.50	3.60	6.00	7.50	6.00

Unequal variance condition [where Population #1 has smaller variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 2.5$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 3.25$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 3.7$
$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/2.5$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/3.25$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/3.7$

Unequal variance condition [where Population #1 has larger variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/2.5$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/1.75$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/1.3$
$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 2.5$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 1.75$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 1.3$

---

d^2=0.754

Equal variance condition:

	X1	X2	X3	X4	X5	X6	X7	X8
$\sigma^2_{(1)}=\sigma^2_{(2)}=\sigma^2_{\text{(pooled)}}$	9.00	18.00	18.00	13.50	10.8	18.00	22.5	18.00

Unequal variance condition [where Population #1 has smaller variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 2.5$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 3.25$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} / 3.7$
$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/2.5$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/3.25$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} \times 4/3.7$

Unequal variance condition [where Population #1 has larger variances]:

Population proportions: .50:.50	Population proportions: .25:.75	Population proportions: .10:.90
$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/2.5$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/1.75$	$\sigma^2_{(1)} = \sigma^2_{\text{(pooled)}} \times 4/1.3$
$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 2.5$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 1.75$	$\sigma^2_{(2)} = \sigma^2_{\text{(pooled)}} / 1.3$

Note. d^2 is group separation measured in Mahalanobis distance: $(\mu_1 - \mu_2)'\Sigma_{pooled}^{-1}(\mu_1 - \mu_2)$, where $\Sigma_{pooled} = \sigma_{pooled} \times R \times \sigma_{pooled}$, and σ_{pooled} is a diagonal matrix with pooled standard deviations of the variables in the diagonal.

Table G1-1

Means and Standard Deviations of Group 1 Error Rates of 200 Replications for Data Pattern I

Covariance matrices = Equal and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.536	.521	.652	.660	.800	.817
		Std	.131	.091	.136	.082	.119	.082
	LDF	Mean	.298	.302	.337	.340	.379	.369
		Std	.114	.077	.118	.076	.131	.080
	LR	Mean	.292	.299	.332	.337	.371	.363
		Std	.121	.078	.118	.077	.128	.082
	KM	Mean	.031	.029	.027	.022	.022	.022
		Std	.0041	.026	.040	.022	.033	.026
0.25 : 0.75	LPM	Mean	.183	.174	.222	.218	.283	.293
		Std	.061	.039	.069	.048	.076	.052
	LDF	Mean	.174	.165	.200	.193	.229	.233
		Std	.059	.038	.066	.045	.067	.049
	LR	Mean	.171	.165	.201	.194	.229	.232
		Std	.061	.037	.068	.046	.068	.050
	KM	Mean	.069	.068	.069	.064	.064	.059
		Std	.038	.028	.036	.026	.039	.024
0.50 : 0.50	LPM	Mean	.0074	.070	.101	.100	.129	.098
		Std	.029	.019	.032	.023	.035	.022
	LDF	Mean	.074	.071	.101	.100	.129	.098
		Std	.029	.019	.032	.023	.035	.022
	LR	Mean	.074	.072	.102	.100	.129	.098
		Std	.029	.019	.035	.025	.037	.023
	KM	Mean	.103	.101	.100	.099	.093	.099
		Std	.035	.027	.033	.025	.030	.023

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.930	.945	.979	.983	.996	.998
		Std	.070	.049	.039	.025	.017	.008
	LDF	Mean	.726	.721	.772	.773	.820	.803
		Std	.116	.083	.115	.072	.115	.080
	LR	Mean	.712	.714	.760	.767	.812	.798
		Std	.120	.088	.121	.077	.119	.082
	KM	Mean	.094	.111	.117	.106	.110	.094
		Std	.076	.054	.070	.046	.070	.050
0.25 : 0.75	LPM	Mean	.463	.454	.553	.545	.692	.698
		Std	.093	.058	.091	.053	.091	.057
	LDF	Mean	.438	.426	.495	.485	.573	.582
		Std	.091	.056	.085	.051	.087	.060
	LR	Mean	.436	.425	.493	.484	.570	.579
		Std	.092	.057	.084	.052	.087	.061
	KM	Mean	.170	.166	.151	.161	.154	.146
		Std	.060	.041	.051	.044	.054	.039
0.50 : 0.50	LPM	Mean	.154	.159	.229	.228	.320	.319
		Std	.037	.031	.049	.035	.055	.038
	LDF	Mean	.155	.161	.229	.228	.319	.317
		Std	.038	.031	.049	.035	.055	.038
	LR	Mean	.156	.161	.230	.228	.318	.317
		Std	.039	.031	.049	.036	.054	.038
	KM	Mean	.245	.248	.226	.226	.212	.217
		Std	.051	.037	.050	.036	.052	.033

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.999	.999	1.00	1.00	1.00	1.00
		Std	.008	.005	.000	.002	.000	.000
	LDF	Mean	.958	.960	.974	.973	.981	.983
		Std	.054	.040	.043	.034	.034	.027
	LR	Mean	.953	.958	.097	.971	.979	.982
		Std	.057	.043	.048	.035	.038	.027
	KM	Mean	.238	.223	.234	.239	.221	.223
		Std	.098	.063	.108	.069	.099	.068
0.25 : 0.75	LPM	Mean	.738	.751	.861	.853	.950	.950
		Std	.084	.056	.075	.057	.048	.034
	LDF	Mean	.702	.713	.808	.787	.879	.872
		Std	.086	.054	.081	.063	.073	.056
	LR	Mean	.701	.712	.803	.787	.877	.870
		Std	.084	.054	.083	.063	.073	.057
	KM	Mean	.288	.280	.278	.272	.269	.263
		Std	.064	.046	.064	.051	.069	.051
0.50 : 0.50	LPM	Mean	.188	.185	.339	.334	.522	.525
		Std	.050	.033	.053	.042	.069	.046
	LDF	Mean	.190	.189	.339	.334	.518	.521
		Std	.050	.034	.053	.042	.069	.046
	LR	Mean	.191	.189	.339	.334	.518	.520
		Std	.050	.034	.053	.042	.070	.046
	KM	Mean	.350	.355	.337	.332	.318	.318
		Std	.064	.044	.055	.041	.062	.040

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.520	.559	.772	.782	.934	.954
		Std	.171	.120	.137	.100	.080	.046
	LDF	Mean	.145	.154	.206	.201	.282	.255
		Std	.092	.068	.115	.073	.125	.085
	LR	Mean	.170	.177	.215	.209	.264	.240
		Std	.093	.072	.116	.074	.128	.079
	KM	Mean	.000	.000	.000	.000	.000	.000
		Std	.000	.001	.000	.002	.000	.000
0.25 : 0.75	LPM	Mean	.047	.042	.076	.079	.175	.160
		Std	.035	.023	.043	.031	.073	.049
	LDF	Mean	.042	.037	.058	.059	.101	.088
		Std	.034	.021	.035	.026	.049	.036
	LR	Mean	.092	.085	.017	.107	.151	.136
		Std	.051	.033	.050	.039	.057	.044
	KM	Mean	.001	.002	.001	.002	.001	.001
		Std	.006	.004	.004	.004	.003	.003
0.50 : 0.50	LPM	Mean	.012	.012	.022	.020	.040	.040
		Std	.013	.007	.017	.010	.022	.017
	LDF	Mean	.013	.012	.022	.020	.040	.039
		Std	.013	.007	.017	.010	.022	.016
	LR	Mean	.043	.039	.064	.060	.092	.088
		Std	.023	.015	.028	.020	.036	.027
	KM	Mean	.007	.006	.006	.005	.006	.006
		Std	.009	.006	.009	.005	.007	.006

Covariance matrices = 1:4 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.996	.999	1.00	1.00	1.00	1.00
		Std	.016	.004	.006	.002	.000	.000
	LDF	Mean	.851	.850	.905	.911	.949	.956
		Std	.102	.073	.089	.057	.064	.042
	LR	Mean	.759	.773	.825	.838	.889	.902
		Std	.129	.093	.125	.086	.102	.068
	KM	Mean	.008	.006	.007	.006	.004	.006
		Std	.021	.013	.020	.012	.014	.012
0.25 : 0.75	LPM	Mean	.380	.403	.610	.597	.816	.836
		Std	.101	.076	.126	.089	.100	.070
	LDF	Mean	.340	.357	.502	.487	.637	.653
		Std	.093	.071	.123	.082	.115	.088
	LR	Mean	.351	.370	.476	.464	.585	.594
		Std	.088	.065	.112	.076	.108	.086
	KM	Mean	.020	.017	.016	.015	.015	.015
		Std	.020	.013	.018	.012	.018	.013
0.50 : 0.50	LPM	Mean	.061	.059	.126	.117	.241	.242
		Std	.028	.017	.039	.026	.059	.044
	LDF	Mean	.061	.060	.126	.117	.239	.238
		Std	.029	.017	.039	.026	.058	.044
	LR	Mean	.091	.088	.167	.157	.275	.273
		Std	.035	.021	.044	.030	.059	.040
	KM	Mean	.046	.044	.040	.038	.037	.036
		Std	.030	.017	.024	.017	.022	.016

Covariance matrices = 1:4 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	1.00	1.00	1.00	1.00	1.00	1.00
		Std	.000	.000	.000	.000	.000	.000
	LDF	Mean	.999	.999	1.00	1.00	1.00	1.00
		Std	.007	.005	.005	.002	.003	.000
	LR	Mean	.996	.997	.999	.999	.999	1.00
		Std	.015	.009	.007	.005	.010	.002
	KM	Mean	.063	.059	.065	.061	.060	.055
		Std	.054	.036	.060	.041	.056	.040
0.25 : 0.75	LPM	Mean	.849	.859	.963	.970	.998	.998
		Std	.095	.061	.043	.029	.007	.008
	LDF	Mean	.807	.812	.920	.929	.978	.983
		Std	.101	.070	.067	.048	.033	.021
	LR	Mean	.777	.783	.892	.905	.961	.971
		Std	.105	.073	.079	.058	.047	.029
	KM	Mean	.084	.081	.083	.080	.077	.077
		Std	.045	.034	.049	.032	.047	.032
0.50 : 0.50	LPM	Mean	.093	.089	.246	.250	.553	.560
		Std	.035	.024	.056	.039	.094	.074
	LDF	Mean	.095	.092	.246	.250	.548	.552
		Std	.035	.025	.056	.039	.093	.073
	LR	Mean	.109	.104	.265	.268	.544	.549
		Std	.038	.027	.056	.040	.088	.069
	KM	Mean	.153	.142	.133	.126	.119	.116
		Std	.056	.041	.054	.038	.047	.034

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.545	.543	.598	.593	.674	.678
		Std	.117	.090	.115	.084	.126	.090
	LDF	Mean	.412	.401	.404	.405	.430	.426
		Std	.111	.085	.117	.080	.123	.082
	LR	Mean	.397	.396	.401	.407	.429	.447
		Std	.115	.091	.122	.085	.129	.092
	KM	Mean	.218	.205	.193	.184	.170	.163
		Std	.100	.072	.105	.071	.103	.068
0.25 : 0.75	LPM	Mean	.279	.276	.308	.296	.344	.335
		Std	.068	.051	.070	.048	.072	.050
	LDF	Mean	.271	.266	.285	.276	.302	.297
		Std	.067	.050	.069	.047	.068	.047
	LR	Mean	.218	.211	.238	.228	.259	.252
		Std	.063	.046	.065	.044	.067	.046
	KM	Mean	.250	.250	.252	.243	.243	.243
		Std	.065	.048	.069	.046	.067	.045
0.50 : 0.50	LPM	Mean	.123	.125	.160	.151	.187	.180
		Std	.036	.026	.039	.027	.040	.030
	LDF	Mean	.124	.126	.160	.151	.187	.179
		Std	.037	.027	.039	.027	.040	.030
	LR	Mean	.086	.087	.111	.101	.128	.123
		Std	.031	.021	.034	.022	.036	.026
	KM	Mean	.227	.229	.224	.215	.219	.210
		Std	.047	.035	.046	.034	.043	.034

Covariance matrices = 4:1 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.827	.844	.877	.887	.926	.936
		Std	.109	.076	.101	.069	.077	.056
	LDF	Mean	.641	.664	.672	.672	.700	.678
		Std	.121	.085	.125	.082	.126	.093
	LR	Mean	.674	.704	.710	.721	.736	.731
		Std	.127	.085	.127	.085	.129	.090
	KM	Mean	.278	.276	.262	.272	.266	.255
		Std	.101	.075	.105	.063	.104	.065
0.25 : 0.75	LPM	Mean	.480	.489	.546	.541	.600	.607
		Std	.078	.053	.079	.060	.087	.060
	LDF	Mean	.463	.470	.507	.503	.533	.530
		Std	.077	.051	.077	.056	.083	.059
	LR	Mean	.443	.449	.497	.492	.536	.533
		Std	.078	.052	.082	.055	.089	.062
	KM	Mean	.358	.368	.348	.348	.327	.322
		Std	.073	.056	.076	.053	.075	.055
0.50 : 0.50	LPM	Mean	.211	.209	.286	.281	.348	.347
		Std	.050	.030	.048	.034	.053	.041
	LDF	Mean	.212	.210	.286	.281	.347	.345
		Std	.050	.030	.048	.034	.053	.041
	LR	Mean	.195	.195	.254	.250	.310	.309
		Std	.045	.027	.044	.032	.052	.040
	KM	Mean	.395	.396	.387	.389	.369	.376
		Std	.059	.037	.058	.038	.055	.041

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.962	.976	.975	.986	.984	.995
		Std	.055	.034	.048	.025	.035	.014
	LDF	Mean	.852	.857	.858	.872	.864	.890
		Std	.112	.087	.112	.082	.109	.077
	LR	Mean	.875	.882	.879	.902	.890	.918
		Std	.106	.077	.107	.072	.098	.066
	KM	Mean	.364	.352	.346	.367	.361	.347
		Std	.112	.077	.102	.080	.122	.079
0.25 : 0.75	LPM	Mean	.698	.695	.753	.767	.838	.843
		Std	.097	.059	.089	.068	.089	.064
	LDF	Mean	.669	.667	.704	.712	.757	.755
		Std	.095	.056	.089	.066	.099	.066
	LR	Mean	.668	.667	.711	.720	.772	.774
		Std	.098	.058	.091	.068	.098	.066
	KM	Mean	.400	.411	.391	.391	.380	.379
		Std	.071	.054	.071	.053	.071	.052
0.50 : 0.50	LPM	Mean	.235	.227	.376	.371	.501	.501
		Std	.051	.040	.048	.038	.060	.038
	LDF	Mean	.237	.230	.376	.371	.499	.498
		Std	.051	.040	.048	.038	.060	.038
	LR	Mean	.238	.232	.364	.360	.484	.483
		Std	.050	.039	.048	.038	.060	.037
	KM	Mean	.464	.474	.458	.459	.444	.445
		Std	.058	.040	.059	.043	.063	.045

Table G1-2 *Means and Standard Deviations of Group 1 Error Rates of 200 Replications for Data Pattern II*

Covariance matrices = Equal and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.526	.530	.647	.652	.791	.801
		Std	.135	.101	.135	.091	.116	.076
	LDF	Mean	.308	.299	.340	.336	.376	.362
		Std	.114	.075	.114	.083	.124	.082
	LR	Mean	.309	.300	.337	.330	.361	.355
		Std	.119	.080	.122	.085	.134	.083
	KM	Mean	.475	.465	.442	.475	.476	.463
		Std	.143	.111	.136	.107	.139	.106
0.25 : 0.75	LPM	Mean	.185	.179	.226	.210	.291	.282
		Std	.062	.040	.061	.046	.074	.056
	LDF	Mean	.176	.170	.203	.186	.236	.223
		Std	.060	.040	.057	.045	.067	.050
	LR	Mean	.179	.172	.206	.187	.240	.222
		Std	.064	.041	.061	.046	.068	.050
	KM	Mean	.429	.441	.422	.455	.445	.443
		Std	.122	.092	.112	.086	.110	.093
0.50 : 0.50	LPM	Mean	.080	.076	.107	.098	.137	.132
		Std	.027	.021	.034	.026	.040	.026
	LDF	Mean	.081	.076	.107	.098	.136	.132
		Std	.027	.021	.034	.026	.040	.026
	LR	Mean	.082	.078	.110	.099	.139	.131
		Std	.030	.022	.036	.026	.043	.028
	KM	Mean	.402	.423	.392	.406	.401	.414
		Std	.105	.085	.117	.104	.109	.096

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.926	.942	.969	.972	.990	.995
		Std	.066	.047	.047	.031	.027	.012
	LDF	Mean	.722	.719	.749	.756	.807	.807
		Std	.117	.085	.117	.088	.108	.076
	LR	Mean	.706	.708	.726	.746	.789	.793
		Std	.117	.092	.123	.087	.111	.085
	KM	Mean	.496	.489	.497	.485	.487	.499
		Std	.121	.101	.129	.087	.124	.090
0.25 : 0.75	LPM	Mean	.459	.450	.552	.556	.679	.674
		Std	.079	.056	.081	.065	.097	.062
	LDF	Mean	.435	.426	.501	.497	.573	.561
		Std	.079	.055	.081	.059	.093	.064
	LR	Mean	.428	.425	.494	.495	.565	.556
		Std	.079	.056	.082	.059	.096	.065
	KM	Mean	.476	.493	.482	.485	.494	.482
		Std	.083	.063	.084	.065	.088	.057
0.50 : 0.50	LPM	Mean	.157	.161	.243	.234	.327	.331
		Std	.042	.030	.043	.032	.057	.037
	LDF	Mean	.158	.163	.243	.234	.325	.329
		Std	.042	.030	.043	.032	.057	.036
	LR	Mean	.161	.163	.242	.233	.326	.329
		Std	.043	.031	.045	.032	.057	.037
	KM	Mean	.486	.492	.490	.490	.488	.490
		Std	.062	.049	.062	.046	.067	.052

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.995	.998	.999	1.00	.999	1.00
		Std	.017	.006	.008	.003	.005	.000
	LDF	Mean	.937	.945	.955	.966	.965	.983
		Std	.069	.042	.056	.036	.046	.024
	LR	Mean	.929	.943	.948	.963	.955	.979
		Std	.073	.042	.061	.038	.056	.026
	KM	Mean	.512	.499	.493	.501	.484	.503
		Std	.113	.078	.133	.080	.108	.086
0.25 : 0.75	LPM	Mean	.722	.735	.831	.843	.926	.940
		Std	.083	.064	.072	.053	.050	.038
	LDF	Mean	.688	.700	.778	.779	.849	.859
		Std	.085	.066	.081	.058	.066	.057
	LR	Mean	.684	.698	.774	.776	.842	.855
		Std	.086	.068	.081	.060	.069	.058
	KM	Mean	.501	.496	.505	.495	.489	.498
		Std	.083	.056	.074	.055	.078	.054
0.50 : 0.50	LPM	Mean	.197	.191	.346	.341	.516	.512
		Std	.046	.033	.057	.037	.063	.047
	LDF	Mean	.200	.194	.346	.341	.513	.509
		Std	.046	.033	.057	.037	.063	.047
	LR	Mean	.201	.194	.348	.341	.513	.507
		Std	.046	.034	.057	.038	.063	.047
	KM	Mean	.498	.500	.495	.498	.493	.495
		Std	.062	.042	.062	.043	.059	.044

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.497	.537	.741	.754	.932	.941
		Std	.146	.119	.148	.101	.080	.053
	LDF	Mean	.149	.150	.195	.186	.266	.255
		Std	.088	.065	.099	.077	.126	.089
	LR	Mean	.179	.182	.218	.195	.255	.244
		Std	.103	.077	.109	.077	.130	.082
	KM	Mean	.318	.369	.359	.364	.393	.378
		Std	.166	.129	.171	.127	.187	.126
0.25 : 0.75	LPM	Mean	.049	.043	.074	.079	.156	.161
		Std	.032	.021	.047	.032	.064	.050
	LDF	Mean	.044	.036	.056	.058	.089	.094
		Std	.030	.019	.040	.025	.047	.036
	LR	Mean	.100	.089	.111	.112	.143	.143
		Std	.051	.035	.058	.037	.056	.041
	KM	Mean	.192	.225	.215	.235	.253	.315
		Std	.141	.125	.124	.108	.157	.134
0.50 : 0.50	LPM	Mean	.014	.011	.022	.022	.040	.041
		Std	.013	.008	.017	.011	.020	.016
	LDF	Mean	.014	.011	.022	.022	.040	.040
		Std	.013	.008	.017	.011	.020	.016
	LR	Mean	.046	.041	.069	.061	.092	.091
		Std	.026	.016	.028	.020	.036	.025
	KM	Mean	.150	.176	.165	.181	.192	.214
		Std	.132	.121	.127	.119	.128	.123

Covariance matrices = 1:4 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.992	.997	1.00	1.00	1.00	1.00
		Std	.026	.011	.004	.000	.000	.000
	LDF	Mean	.820	.843	.881	.900	.930	.945
		Std	.114	.079	.096	.067	.076	.045
	LR	Mean	.723	.749	.794	.813	.840	.880
		Std	.131	.096	.124	.098	.129	.076
	KM	Mean	.453	.428	.453	.454	.457	.455
		Std	.160	.107	.146	.121	.151	.118
0.25 : 0.75	LPM	Mean	.385	.393	.560	.570	.805	.821
		Std	.098	.071	.111	.086	.095	.067
	LDF	Mean	.348	.349	.467	.465	.626	.639
		Std	.095	.065	.102	.080	.111	.080
	LR	Mean	.361	.361	.446	.450	.564	.575
		Std	.089	.061	.093	.073	.107	.078
	KM	Mean	.398	.416	.413	.426	.416	.422
		Std	.126	.093	.123	.105	.127	.094
0.50 : 0.50	LPM	Mean	.063	.060	.120	.116	.250	.242
		Std	.029	.019	.038	.027	.062	.041
	LDF	Mean	.064	.061	.120	.116	.247	.239
		Std	.029	.019	.038	.027	.062	.041
	LR	Mean	.097	.092	.165	.158	.285	.277
		Std	.034	.024	.045	.030	.060	.040
	KM	Mean	.384	.424	.391	.412	.398	.422
		Std	.120	.084	.120	.090	.109	.080

Covariance matrices = 1:4 and Group separation = 0.7								
Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	1.00	1.00	1.00	1.00	1.00	1.00
		Std	.000	.000	.000	.000	.000	.000
	LDF	Mean	.998	.999	1.00	.999	1.00	1.00
		Std	.013	.004	.002	.005	.004	.000
	LR	Mean	.991	.998	.999	.998	.999	1.00
		Std	.028	.008	.006	.008	.009	.003
	KM	Mean	.468	.473	.446	.476	.486	.449
		Std	.143	.100	.139	.097	.149	.111
0.25 : 0.75	LPM	Mean	.815	.836	.943	.961	.994	.997
		Std	.096	.065	.051	.033	.014	.006
	LDF	Mean	.773	.787	.891	.915	.961	.974
		Std	.102	.073	.073	.051	.042	.026
	LR	Mean	.146	.757	.856	.886	.936	.958
		Std	.102	.077	.082	.062	.060	.036
	KM	Mean	.439	.462	.447	.460	.451	.470
		Std	.110	.077	.108	.084	.112	.083
0.50 : 0.50	LPM	Mean	.101	.094	.245	.248	.529	.551
		Std	.038	.022	.058	.037	.090	.072
	LDF	Mean	.103	.096	.245	.248	.524	.543
		Std	.038	.022	.058	.037	.090	.072
	LR	Mean	.122	.110	.267	.267	.523	.541
		Std	.041	.024	.057	.037	.087	.068
	KM	Mean	.444	.448	.444	.445	.443	.459
		Std	.078	.062	.086	.066	.095	.067

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.538	.541	.607	.599	.708	.676
		Std	.123	.081	.112	.093	.109	.085
	LDF	Mean	.421	.411	.438	.419	.483	.434
		Std	.115	.080	.103	.085	.123	.083
	LR	Mean	.402	.401	.436	.426	.480	.442
		Std	.113	.085	.114	.092	.127	.087
	KM	Mean	.502	.488	.485	.501	.500	.493
		Std	.126	.091	.119	.083	.126	.090
0.25 : 0.75	LPM	Mean	.286	.279	.308	.310	.350	.347
		Std	.071	.046	.072	.051	.077	.050
	LDF	Mean	.278	.269	.291	.293	.315	.309
		Std	.072	.045	.068	.050	.073	.049
	LR	Mean	.226	.215	.241	.239	.272	.268
		Std	.069	.044	.068	.046	.072	.046
	KM	Mean	.490	.492	.467	.481	.467	.484
		Std	.094	.068	.092	.069	.095	.064
0.50 : 0.50	LPM	Mean	.133	.127	.160	.159	.188	.187
		Std	.035	.025	.041	.028	.042	.032
	LDF	Mean	.133	.128	.160	.159	.188	.186
		Std	.035	.025	.041	.028	.042	.032
	LR	Mean	.096	.090	.116	.109	.132	.128
		Std	.032	.021	.037	.023	.037	.027
	KM	Mean	.484	.490	.480	.483	.450	.451
		Std	.079	.057	.080	.063	.078	.074

Covariance matrices = 4:1 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.821	.841	.856	.871	.907	.922
		Std	.105	.069	.104	.067	.082	.055
	LDF	Mean	.670	.668	.674	.673	.698	.686
		Std	.121	.076	.119	.088	.115	.092
	LR	Mean	.685	.696	.696	.707	.713	.726
		Std	.126	.080	.127	.088	.122	.096
	KM	Mean	.501	.510	.513	.509	.509	.497
		Std	.110	.081	.117	.090	.113	.076
0.25 : 0.75	LPM	Mean	.498	.495	.539	.539	.602	.605
		Std	.072	.050	.078	.052	.082	.056
	LDF	Mean	.479	.478	.505	.504	.540	.535
		Std	.072	.051	.076	.050	.081	.056
	LR	Mean	.475	.457	.488	.490	.530	.535
		Std	.072	.050	.080	.051	.084	.057
	KM	Mean	.514	.508	.515	.510	.519	.511
		Std	.080	.050	.079	.053	.075	.051
0.50 : 0.50	LPM	Mean	.215	.211	.298	.287	.357	.355
		Std	.046	.033	.047	.034	.056	.035
	LDF	Mean	.216	.212	.298	.287	.356	.353
		Std	.046	.033	.047	.034	.056	.035
	LR	Mean	.201	.196	.265	.256	.315	.315
		Std	.044	.030	.046	.033	.054	.035
	KM	Mean	.519	.518	.519	.516	.514	.504
		Std	.061	.042	.059	.043	.059	.045

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.929	.953	.954	.974	.971	.990
		Std	.067	.050	.058	.032	.045	.020
	LDF	Mean	.809	.837	.815	.842	.825	.865
		Std	.101	.078	.102	.072	.101	.076
	LR	Mean	.826	.861	.831	.870	.841	.891
		Std	.106	.073	.101	.069	.094	.071
	KM	Mean	.499	.509	.512	.505	.515	.503
		Std	.129	.079	.112	.079	.114	.084
0.25 : 0.75	LPM	Mean	.668	.686	.741	.751	.798	.823
		Std	.079	.060	.078	.066	.087	.063
	LDF	Mean	.646	.658	.696	.700	.726	.741
		Std	.080	.060	.079	.066	.089	.068
	LR	Mean	.642	.656	.698	.708	.735	.758
		Std	.082	.063	.082	.068	.092	.067
	KM	Mean	.519	.508	.520	.506	.512	.508
		Std	.075	.054	.076	.047	.075	.049
0.50 : 0.50	LPM	Mean	.254	.243	.390	.382	.511	.505
		Std	.050	.040	.053	.035	.057	.037
	LDF	Mean	.256	.246	.390	.382	.509	.502
		Std	.050	.040	.053	.035	.057	.037
	LR	Mean	.255	.247	.379	.369	.491	.488
		Std	.049	.039	.051	.034	.059	.038
	KM	Mean	.522	.516	.525	.515	.520	.512
		Std	.057	.037	.055	.040	.054	.040

Table G1-3

ANOVA Results, Eta-squared, and Partial-Omega squared of Group 1 Error Rate for Data Pattern I on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	3104.29	1552.15	103075.0	<.0001	.2315	.6140
Equality of covariance (COV)	2	139.17	69.59	4621.06	<.0001	.0104	.0666
Group separation (GS)	2	3335.82	1667.91	110763.0	<.0001	.2487	.6309
Sample representativeness (SR)	2	196.25	98.13	6516.39	<.0001	.0146	.0914
Sample size (SS)	1	0.0035	0.0035	0.23	.6289	<.0001	<.0001
PP*COV	4	55.37	13.84	919.30	<.0001	.0041	.0276
PP*GS	4	327.65	81.91	5439.72	<.0001	.0244	.1437
PP*SR	4	23.48	5.87	389.84	<.0001	.0018	.0119
PP*SS	2	0.12	0.058	3.85	.0213	<.0001	<.0001
COV*GS	4	54.26	13.56	90.80	<.0001	.0040	.0270
COV*SR	4	13.43	3.36	222.98	<.0001	.0010	.0068
COV*SS	2	0.051	0.026	1.70	.1819	<.0001	<.0001
GS*SR	4	16.02	4.00	265.93	<.0001	.0012	.0081
GS*SS	2	0.14	0.071	4.72	.0090	<.0001	.0001
SR*SS	2	0.049	0.025	1.63	.1956	<.0001	<.0001
Error (between)	32358	487.26	0.015				
Within-subject effects:							
Method (4M)	3	2558.71	852.90	279409.0	<.0001	.1908	.8661
4M*PP	6	1616.77	269.46	88274.90	<.0001	.1205	.8034
4M*COV	6	291.08	48.51	15892.80	<.0001	.0217	.4239
4M*GS	6	424.10	7.68	23155.50	<.0001	.0316	.5174
4M*SR	6	91.93	15.32	5019.11	<.0001	.0069	.1885
4M*SS	3	0.052	0.017	5.69	.0007	<.0001	.0001
4M*PP*COV	12	2.08	1.67	548.27	<.0001	.0015	.0482
4M*PP*GS	12	25.92	2.91	685.15	<.0001	.0187	.3881
4M*PP*SR	12	11.64	0.97	317.75	<.0001	.0009	.0285
4M*PP*SS	6	0.12	0.021	6.77	<.0001	<.0001	.0003
4M*COV*GS	12	83.92	6.99	2291.08	<.0001	.0063	.1749
4M*COV*SR	12	2.58	0.22	7.45	<.0001	.0002	.0064
4M*COV*SS	6	0.022	0,0037	1.20	.3047	<.0001	<.0001
4M*GS*SR	12	1.64	0.89	29.48	<.0001	.0008	.0261
4M*GS*SS	6	0.045	0.0075	2.44	.0230	<.0001	.0001
4M*SR*SS	6	0.016	0.0027	0.88	.5095	<.0001	<.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	296.32	0.0031				

Note. Eta square (η^2) is defined as: $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \frac{df_{effect}(F_{effect}-1)}{df_{effect}(F_{effect}-1)+N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table G1-4

ANOVA Results, Eta-squared, and Partial-Omega squared of Group 1 Error Rate for Data Pattern II on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	3402.68	1701.34	112308.0	<.0001	.3286	.6341
Equality of covariance (COV)	2	53.01	26.51	1749.78	<.0001	.0051	.0263
Group separation (GS)	2	2928.04	1464.02	96642.30	<.0001	.2828	.5986
Sample representativeness (SR)	2	21.15	105.08	6936.18	<.0001	.0203	.0967
Sample size (SS)	1	0.25	0.25	16.72	<.0001	<.0001	.0001
PP*COV	4	67.36	16.84	1111.67	<.0001	.0065	.0331
PP*GS	4	296.40	74.10	4891.46	<.0001	.0286	.1311
PP*SR	4	22.13	5.53	365.23	<.0001	.0021	.0111
PP*SS	2	0.22	0.11	7.39	.0006	<.0001	.0001
COV*GS	4	148.57	37.14	2451.86	<.0001	.0143	.0703
COV*SR	4	14.86	3.72	245.24	<.0001	.0014	.0075
COV*SS	2	0.32	0.16	1.45	<.0001	<.0001	.0001
GS*SR	4	13.71	3.43	226.30	<.0001	.0013	.0069
GS*SS	2	0.34	0.17	11.07	<.0001	<.0001	.0002
SR*SS	2	0.004	0.002	0.13	.8769	<.0001	<.0001
Error (between)	32358	49.19	.015				
Within-subject effects:							
Method (4M)	3	173.36	57.79	11165.10	<.0001	.0167	.2054
4M*PP	6	1085.50	18.92	34955.60	<.0001	.1048	.6181
4M*COV	6	45.41	7.57	1462.22	<.0001	.0044	.0634
4M*GS	6	547.84	91.31	17641.70	<.0001	.0529	.4496
4M*SR	6	68.16	11.36	2194.91	<.0001	.0066	.0922
4M*SS	3	0.16	0.052	1.01	<.0001	<.0001	.0002
4M*PP*COV	12	15.13	1.26	243.58	<.0001	.0015	.0220
4M*PP*GS	12	225.40	18.78	3629.16	<.0001	.0218	.2515
4M*PP*SR	12	11.34	0.94	182.52	<.0001	.0011	.0165
4M*PP*SS	6	0.28	0.046	8.87	<.0001	<.0001	.0004
4M*COV*GS	12	18.43	1.54	296.70	<.0001	.0018	.0266
4M*COV*SR	12	1.51	0.13	24.28	<.0001	.0001	.0022
4M*COV*SS	6	0.28	0.046	8.95	<.0001	<.0001	.0004
4M*GS*SR	12	1.34	0.86	166.41	<.0001	.0010	.0151
4M*GS*SS	6	0.77	0.13	24.90	<.0001	.0001	.0011
4M*SR*SS	6	0.025	0.0041	0.79	.5784	<.0001	<.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	502.42	0.0052				

Note. Eta square (η^2) is defined as: $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table G1-5

Means and Standard Deviations of Group 1 Error Rate for Method by Population Proportion Interaction

Population proportions	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Data Pattern I								
0.10:0.90	.873	.180	.682	.292	.680	.289	.138	.138
0.25:0.75	.545	.291	.495	.278	.489	.271	.176	.145
0.50:0.50	.216	.153	.216	.152	.216	.145	.212	.154
Data Pattern II								
0.10:0.90	.866	.180	.676	.285	.670	.281	.473	.127
0.25:0.75	.537	.284	.488	.270	.482	.260	.448	.123
0.50:0.50	.221	.152	.220	.151	.221	.143	.437	.128

Table G1-6

Means and Standard Deviations of Group 1 Error Rate for Method by Equality of Covariance Matrices Interaction

Equality of covariance matrices	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
1:4	.540	.402	.449	.384	.451	.357	.038	.051
Equal	.551	.337	.472	.303	.470	.301	.169	.111
4:1	.543	.285	.472	.238	.464	.262	.318	.110
				Data Pattern II				
1:4	.533	.399	.442	.379	.444	.348	.381	.155
Equal	.549	.331	.470	.294	.466	.291	.475	.096
4:1	.542	.276	.474	.227	.463	.248	.502	.081

Table G1-7

Means and Standard Deviations of Group 1 Error Rate for Method by Group Separation Interaction

Group separation	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
6.7	.326	.281	.198	.141	.197	.132	.095	.103
2.2	.581	.319	.500	.258	.492	.247	.173	.144
0.7	.728	.305	.695	.288	.696	.287	.258	.148
				Data Pattern II				
6.7	.325	.276	.201	.144	.202	.134	.394	.157
2.2	.578	.314	.498	.253	.486	.237	.475	.102
0.7	.721	.299	.685	.280	.684	.277	.490	.089

Table G1-8

Means and Standard Deviations of Group 1 Error Rate for Method by Sample Representativeness Interaction

Sample representativeness	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
20% over	.468	.347	.406	.316	.404	.309	.182	.152
Equal	.542	.340	.462	.311	.459	.306	.176	.149
20% under	.625	.328	.525	.303	.522	.301	.168	.145
				Data Pattern II				
20% over	.465	.341	.405	.308	.402	.300	.451	.131
Equal	.537	.335	.458	.304	.455	.296	.451	.127
20% under	.622	.323	.522	.296	.516	.291	.456	.122

Table G1-9

Means and Standard Deviations of Group 1 Error Rate for Method by Population Proportions by Group Separation Interaction

Group separation	Population proportions	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		Data Pattern I							
6.7	0.10:0.90	.675	.175	.319	.135	.319	.134	.072	.100
	0.25:0.75	.211	.111	.182	.106	.182	.077	.104	.111
	0.50:0.50	.091	.062	.091	.062	.089	.038	.109	.092
2.2	0.10:0.90	.951	.078	.781	.138	.768	.122	.126	.125
	0.25:0.75	.573	.147	.499	.113	.488	.104	.174	.143
	0.50:0.50	.219	.097	.219	.097	.221	.083	.218	.147
0.7	0.10:0.90	.993	.025	.945	.084	.953	.074	.215	.145
	0.25:0.75	.852	.118	.803	.123	.797	.119	.249	.140
	0.50:0.50	.339	.162	.338	.160	.338	.154	.308	.144
		Data Pattern II							
6.7	0.10:0.90	.668	.172	.324	.141	.325	.139	.441	.142
	0.25:0.75	.212	.113	.185	.109	.187	.079	.386	.152
	0.50:0.50	.096	.065	.096	.065	.095	.041	.353	.163
2.2	0.10:0.90	.945	.080	.775	.133	.749	.120	.483	.118
	0.25:0.75	.566	.140	.496	.108	.482	.099	.471	.096
	0.50:0.50	.224	.100	.223	.099	.226	.084	.470	.088
0.7	0.10:0.90	.987	.035	.930	.093	.935	.086	.493	.111
	0.25:0.75	.834	.120	.785	.121	.776	.117	.488	.082
	0.50:0.50	.342	.157	.341	.154	.342	.148	.487	.067

Table G1-10

Means and Standard Deviations of Group 1 Error Rate for Method by Equality of Covariance Matrices by Group Separation Interaction

Group separation	Equality of covariance matrices	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		Data Pattern I							
6.7	1:4	.292	.352	.099	.103	.130	.092	.002	.006
	Equal	.329	.264	.211	.123	.209	.122	.063	.043
	4:1	.355	.204	.283	.127	.251	.146	.219	.070
2.2	1:4	.582	.373	.513	.332	.493	.291	.021	.023
	Equal	.591	.314	.501	.237	.498	.235	.164	.073
	4:1	.569	.260	.484	.181	.486	.207	.333	.085
0.7	1:4	.746	.341	.734	.336	.729	.327	.091	.055
	Equal	.733	.297	.705	.282	.703	.281	.280	.081
	4:1	.704	.269	.648	.231	.657	.242	.402	.085
		Data Pattern II							
6.7	1:4	.284	.343	.096	.099	.132	.092	.261	.160
	Equal	.331	.258	.214	.120	.213	.119	.437	.114
	4:1	.361	.205	.295	.134	.262	.151	.483	.089
2.2	1:4	.576	.371	.503	.325	.479	.277	.424	.119
	Equal	.590	.309	.500	.231	.494	.226	.489	.083
	4:1	.568	.252	.491	.180	.485	.201	.511	.075
0.7	1:4	.740	.339	.725	.333	.720	.323	.459	.102
	Equal	.727	.293	.694	.274	.692	.272	.498	.077
	4:1	.696	.256	.636	.212	.641	.222	.513	.075

Table G1-11

Means and Standard Deviations of Group 1 Error Rate for Method by Population Proportion by Equality of Covariance Matrices Interaction

Population Proportion	Equality of covariance matrices	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I					
0.10:0.90	1:4	.918	.165	.704	.362	.681	.349	.022	.040
	Equal	.879	.179	.693	.279	.687	.281	.120	.106
	4:1	.822	.182	.650	.212	.672	.223	.271	.114
0.25:0.75	1:4	.548	.371	.488	.360	.489	.330	.033	.043
	Equal	.549	.274	.498	.258	.496	.257	.167	.098
	4:1	.539	.206	.498	.190	.482	.213	.328	.087
0.50:0.50	1:4	.155	.170	.154	.168	.182	.157	.059	.061
	Equal	.226	.143	.226	.142	.226	.141	.221	.106
	4:1	.268	.121	.268	.121	.240	.129	.354	.110
				Data Pattern II					
0.10:0.90	1:4	.911	.173	.695	.362	.670	.346	.429	.147
	Equal	.874	.180	.685	.275	.677	.274	.486	.113
	4:1	.814	.174	.648	.191	.663	.202	.503	.103
0.25:0.75	1:4	.536	.365	.476	.351	.478	.318	.370	.150
	Equal	.541	.266	.491	.250	.489	.248	.474	.087
	4:1	.535	.194	.498	.179	.479	.201	.502	.073
0.50:0.50	1:4	.154	.165	.153	.163	.184	.153	.344	.157
	Equal	.233	.139	.232	.138	.233	.137	.464	.085
	4:1	.276	.123	.276	.122	.247	.130	.502	.063

Table G2-1

Means and Standard Deviations of Group 2 Error Rates of 200 Replications for Data Pattern I

Covariance matrices = Equal and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.005	.004	.002	.002	.000	.000
		Std	.005	.003	.004	.002	.001	.001
	LDF	Mean	.022	.020	.017	.017	.014	.014
		Std	.013	.008	.011	.008	.010	.007
	LR	Mean	.024	.020	.020	.017	.016	.015
		Std	.014	.009	.012	.008	.012	.008
	KM	Mean	.257	.255	.270	.278	.298	.315
		Std	.060	.041	.057	.043	.058	.044
0.25 : 0.75	LPM	Mean	.049	.050	.037	.035	.025	.023
		Std	.020	.013	.016	.012	.013	.010
	LDF	Mean	.053	.054	.046	.042	.036	.035
		Std	.021	.014	.018	.013	.015	.012
	LR	Mean	.054	.055	.047	.043	.038	.035
		Std	.022	.014	.019	.014	.017	.012
	KM	Mean	.135	.139	.146	.145	.160	.160
		Std	.041	.028	.039	.029	.046	.030
0.50 : 0.50	LPM	Mean	.129	.129	.101	.100	.073	.099
		Std	.035	.025	.034	.023	.028	.023
	LDF	Mean	.129	.129	.101	.100	.074	.099
		Std	.035	.025	.034	.023	.028	.023
	LR	Mean	.131	.128	.104	.102	.075	.099
		Std	.038	.027	.034	.025	.028	.025
	KM	Mean	.092	.093	.100	.101	.106	.099
		Std	.029	.021	.036	.027	.035	.026

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.002	.001	.000	.000	.000	.000
		Std	.004	.002	.001	.001	.000	.000
	LDF	Mean	.020	.020	.016	.015	.011	.010
		Std	.014	.009	.011	.008	.011	.007
	LR	Mean	.022	.021	.018	.016	.012	.010
		Std	.015	.011	.013	.009	.013	.008
	KM	Mean	.389	.405	.403	.408	.415	.416
		Std	.052	.040	.054	.041	.048	.039
0.25 : 0.75	LPM	Mean	.085	.086	.052	.054	.026	.023
		Std	.029	.019	.023	.016	.016	.011
	LDF	Mean	.095	.097	.070	.073	.050	.048
		Std	.030	.020	.025	.018	.023	.015
	LR	Mean	.097	.098	.071	.073	.052	.048
		Std	.031	.020	.025	.018	.023	.015
	KM	Mean	.300	.303	.310	.314	.332	.329
		Std	.051	.034	.052	.039	.051	.039
0.50 : 0.50	LPM	Mean	.324	.320	.232	.232	.154	.154
		Std	.056	.040	.046	.034	.041	.028
	LDF	Mean	.322	.318	.232	.232	.155	.156
		Std	.055	.039	.046	.034	.041	.028
	LR	Mean	.322	.317	.232	.233	.155	.156
		Std	.057	.040	.047	.034	.043	.028
	KM	Mean	.217	.216	.232	.233	.243	.242
		Std	.054	.035	.051	.036	.051	.038

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.001	.000	.000	.000	.000	.000
	LDF	Mean	.007	.005	.004	.003	.003	.002
		Std	.010	.006	.007	.004	.005	.003
	LR	Mean	.008	.005	.005	.004	.004	.002
		Std	.011	.007	.008	.005	.006	.003
	KM	Mean	.443	.447	.454	.455	.458	.459
		Std	.053	.039	.050	.033	.051	.036
0.25 : 0.75	LPM	Mean	.071	.066	.033	.030	.008	.007
		Std	.033	.024	.024	.016	.010	.006
	LDF	Mean	.086	.081	.054	.049	.025	.026
		Std	.036	.026	.030	.021	.020	.014
	LR	Mean	.088	.082	.056	.050	.027	.026
		Std	.036	.026	.031	.021	.021	.014
	KM	Mean	.390	.393	.407	.394	.406	.414
		Std	.052	.035	.051	.036	.054	.036
0.50 : 0.50	LPM	Mean	.529	.522	.335	.337	.189	.182
		Std	.064	.046	.055	.041	.048	.037
	LDF	Mean	.527	.517	.335	.337	.191	.184
		Std	.064	.045	.055	.041	.048	.037
	LR	Mean	.526	.516	.335	.337	.193	.185
		Std	.064	.045	.055	.042	.048	.037
	KM	Mean	.321	.319	.330	.337	.352	.350
		Std	.058	.038	.054	.042	.057	.038

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.007	.006	.003	.003	.001	.000
		Std	.007	.004	.004	.003	.002	.001
	LDF	Mean	.026	.024	.021	.021	.017	.016
		Std	.014	.009	.012	.008	.010	.007
	LR	Mean	.025	.023	.022	.020	.020	.017
		Std	.013	.009	.013	.008	.011	.008
	KM	Mean	.275	.284	.296	.301	.325	.329
		Std	.052	.035	.052	.038	.055	.040
0.25 : 0.75	LPM	Mean	.072	.069	.055	.053	.033	.034
		Std	.023	.016	.021	.015	.016	.011
	LDF	Mean	.076	.073	.064	.062	.048	.050
		Std	.023	.017	.023	.016	.020	.014
	LR	Mean	.052	.050	.046	.045	.039	.039
		Std	.019	.014	.021	.012	.019	.012
	KM	Mean	.204	.202	.213	.210	.228	.230
		Std	.041	.028	.044	.031	.040	.031
0.50 : 0.50	LPM	Mean	.189	.178	.159	.152	.128	.122
		Std	.041	.029	.040	.026	.036	.026
	LDF	Mean	.188	.177	.159	.152	.128	.123
		Std	.041	.029	.040	.026	.036	.026
	LR	Mean	.132	.121	.108	.102	.091	.087
		Std	.037	.025	.034	.023	.031	.022
	KM	Mean	.216	.211	.220	.215	.222	.225
		Std	.042	.033	.042	.031	.045	.037

Covariance matrices = 1:4 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.003	.002	.001	.000	.000	.000
		Std	.005	.003	.002	.001	.001	.000
	LDF	Mean	.027	.027	.020	.018	.012	.011
		Std	.018	.011	.014	.009	.011	.007
	LR	Mean	.040	.038	.033	.027	.021	.019
		Std	.022	.014	.019	.013	.017	.010
	KM	Mean	.434	.427	.437	.434	.434	.441
		Std	.049	.033	.050	.037	.048	.035
0.25 : 0.75	LPM	Mean	.118	.111	.071	.073	.037	.035
		Std	.033	.023	.026	.021	.022	.014
	LDF	Mean	.128	.123	.091	.093	.067	.065
		Std	.033	.024	.029	.023	.026	.018
	LR	Mean	.125	.012	.096	.098	.076	.075
		Std	.031	.022	.028	.022	.026	.018
	KM	Mean	.387	.387	.397	.398	.398	.404
		Std	.048	.037	.049	.034	.050	.031
0.50 : 0.50	LPM	Mean	.355	.353	.286	.285	.208	.206
		Std	.054	.038	.051	.037	.044	.033
	LDF	Mean	.354	.351	.286	.285	.209	.208
		Std	.053	.038	.051	.037	.044	.033
	LR	Mean	.320	.314	.256	.254	.195	.193
		Std	.053	.034	.048	.036	.043	.031
	KM	Mean	.380	.379	.392	.388	.393	.395
		Std	.052	.042	.054	.040	.055	.041

Covariance matrices = 1:4 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.000	.000	.000	.000	.000	.000
	LDF	Mean	.009	.006	.005	.004	.003	.001
		Std	.011	.006	.007	.005	.006	.002
	LR	Mean	.014	.009	.008	.006	.005	.002
		Std	.015	.008	.011	.007	.010	.003
	KM	Mean	.483	.480	.480	.487	.483	.491
		Std	.047	.034	.048	.034	.050	.033
0.25 : 0.75	LPM	Mean	.099	.095	.047	.039	.012	.010
		Std	.036	.026	.028	.018	.014	.009
	LDF	Mean	.116	.113	.069	.063	.035	.033
		Std	.038	.027	.036	.021	.025	.018
	LR	Mean	.127	.123	.082	.074	.045	.042
		Std	.039	.028	.038	.024	.029	.022
	KM	Mean	.471	.477	.473	.473	.475	.478
		Std	.048	.034	.049	.036	.048	.037
0.50 : 0.50	LPM	Mean	.501	.502	.373	.372	.230	.228
		Std	.061	.035	.057	.038	.054	.037
	LDF	Mean	.498	.499	.373	.372	.232	.231
		Std	.061	.035	.057	.038	.054	.037
	LR	Mean	.483	.485	.361	.361	.234	.233
		Std	.061	.037	.056	.037	.052	.036
	KM	Mean	.433	.446	.460	.461	.471	.473
		Std	.060	.044	.065	.044	.060	.043

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.001	.001	.001	.000	.000	.000
		Std	.003	.002	.002	.001	.001	.000
	LDF	Mean	.010	.009	.008	.008	.006	.007
		Std	.009	.005	.008	.006	.006	.005
	LR	Mean	.012	.010	.011	.008	.007	.006
		Std	.011	.006	.010	.006	.008	.005
	KM	Mean	.109	.107	.134	.129	.171	.190
		Std	.072	.050	.082	.071	.095	.081
0.25 : 0.75	LPM	Mean	.016	.013	.010	.008	.005	.006
		Std	.011	.007	.009	.006	.006	.005
	LDF	Mean	.018	.015	.013	.011	.009	.010
		Std	.012	.008	.010	.007	.009	.006
	LR	Mean	.038	.035	.027	.026	.019	.019
		Std	.019	.012	.016	.012	.013	.009
	KM	Mean	.022	.019	.020	.019	.022	.021
		Std	.016	.010	.014	.012	.017	.011
0.50 : 0.50	LPM	Mean	.038	.041	.022	.022	.012	.014
		Std	.020	.015	.016	.012	.012	.009
	LDF	Mean	.038	.040	.022	.022	.012	.014
		Std	.020	.015	.016	.012	.012	.009
	LR	Mean	.090	.091	.063	.062	.043	.042
		Std	.031	.024	.027	.020	.027	.016
	KM	Mean	.005	.005	.005	.006	.007	.007
		Std	.007	.005	.008	.006	.009	.006

Covariance matrices = 4:1 and Group separation = 2.2

Population Proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.001	.000	.000	.000	.000	.000
		Std	.002	.001	.001	.001	.000	.000
	LDF	Mean	.010	.008	.007	.007	.006	.006
		Std	.010	.007	.007	.006	.008	.005
	LR	Mean	.008	.005	.005	.004	.004	.003
		Std	.009	.005	.007	.005	.007	.003
	KM	Mean	.302	.317	.327	.333	.349	.355
		Std	.074	.058	.074	.054	.075	.050
0.25 : 0.75	LPM	Mean	.032	.031	.018	.017	.009	.006
		Std	.017	.012	.013	.009	.009	.005
	LDF	Mean	.038	.038	.028	.027	.020	.018
		Std	.019	.013	.016	.011	.015	.009
	LR	Mean	.049	.048	.031	.031	.020	.017
		Std	.021	.016	.018	.012	.015	.009
	KM	Mean	.110	.106	.127	.125	.145	.143
		Std	.051	.037	.054	.039	.060	.042
0.50 : 0.50	LPM	Mean	.246	.243	.119	.123	.061	.061
		Std	.063	.041	.039	.031	.025	.021
	LDF	Mean	.243	.239	.119	.123	.062	.062
		Std	.062	.041	.039	.031	.026	.021
	LR	Mean	.278	.275	.159	.166	.090	.091
		Std	.062	.042	.045	.034	.032	.025
	KM	Mean	.038	.035	.038	.040	.046	.045
		Std	.022	.017	.026	.018	.025	.018

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.001	.000	.000	.000	.000	.000
	LDF	Mean	.004	.002	.003	.002	.003	.001
		Std	.006	.004	.007	.003	.005	.003
	LR	Mean	.002	.001	.002	.001	.002	.001
		Std	.005	.003	.006	.002	.004	.002
	KM	Mean	.400	.405	.410	.411	.413	.427
		Std	.067	.046	.060	.046	.065	.044
0.25 : 0.75	LPM	Mean	.021	.019	.009	.007	.003	.001
		Std	.017	.012	.011	.007	.007	.003
	LDF	Mean	.029	.027	.019	.016	.012	.009
		Std	.020	.014	.016	.012	.013	.008
	LR	Mean	.031	.027	.018	.014	.009	.006
		Std	.022	.015	.016	.012	.011	.007
	KM	Mean	.253	.249	.272	.268	.296	.302
		Std	.069	.053	.067	.046	.072	.051
0.50 : 0.50	LPM	Mean	.548	.572	.250	.247	.092	.091
		Std	.080	.063	.054	.039	.036	.027
	LDF	Mean	.542	.564	.250	.247	.094	.094
		Std	.088	.063	.054	.039	.036	.027
	LR	Mean	.538	.559	.270	.267	.107	.106
		Std	.084	.060	.052	.040	.042	.030
	KM	Mean	.116	.121	.132	.124	.141	.141
		Std	.049	.035	.051	.038	.057	.043

Table G2-2

Means and Standard Deviations of Group 2 Error Rates of 200 Replications for Data Pattern II

Covariance matrices = Equal and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.006	.005	.002	.002	.001	.000
		Std	.006	.004	.004	.002	.002	.001
	LDF	Mean	.024	.021	.020	.018	.015	.014
		Std	.013	.008	.011	.008	.011	.007
	LR	Mean	.029	.023	.026	.020	.022	.016
		Std	.014	.009	.014	.009	.014	.008
	KM	Mean	.484	.490	.488	.492	.497	.490
		Std	.056	.037	.050	.037	.051	.035
0.25 : 0.75	LPM	Mean	.053	.051	.038	.038	.027	.023
		Std	.020	.014	.019	.011	.016	.009
	LDF	Mean	.056	.055	.045	.045	.038	.035
		Std	.021	.014	.020	.012	.019	.011
	LR	Mean	.059	.056	.050	.047	.043	.037
		Std	.024	.014	.024	.013	.021	.011
	KM	Mean	.453	.468	.457	.480	.477	.476
		Std	.080	.058	.074	.050	.066	.050
0.50 : 0.50	LPM	Mean	.136	.131	.105	.101	.077	.076
		Std	.040	.026	.033	.024	.027	.021
	LDF	Mean	.135	.131	.105	.101	.077	.076
		Std	.040	.026	.033	.024	.028	.021
	LR	Mean	.138	.132	.107	.103	.080	.077
		Std	.040	.028	.035	.025	.029	.021
	KM	Mean	.391	.414	.399	.407	.410	.433
		Std	.112	.095	.126	.104	.111	.088

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.003	.002	.001	.000	.000	.000
		Std	.005	.002	.003	.001	.001	.000
	LDF	Mean	.028	.023	.021	.017	.015	.012
		Std	.018	.010	.015	.008	.012	.007
	LR	Mean	.032	.025	.026	.019	.018	.014
		Std	.020	.011	.016	.010	.014	.008
	KM	Mean	.501	.498	.498	.496	.494	.497
		Std	.052	.033	.048	.036	.055	.033
0.25 : 0.75	LPM	Mean	.090	.090	.062	.056	.032	.027
		Std	.028	.019	.023	.017	.018	.013
	LDF	Mean	.099	.100	.080	.074	.058	.053
		Std	.029	.020	.026	.019	.021	.018
	LR	Mean	.103	.102	.083	.075	.061	.055
		Std	.029	.020	.026	.019	.022	.019
	KM	Mean	.493	.489	.492	.490	.494	.495
		Std	.055	.038	.053	.039	.052	.040
0.50 : 0.50	LPM	Mean	.329	.321	.238	.233	.163	.156
		Std	.048	.043	.046	.033	.044	.029
	LDF	Mean	.328	.319	.238	.233	.165	.158
		Std	.048	.043	.046	.033	.044	.029
	LR	Mean	.325	.318	.239	.235	.166	.158
		Std	.049	.043	.048	.033	.046	.030
	KM	Mean	.492	.487	.492	.489	.487	.490
		Std	.065	.048	.065	.048	.067	.045

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.002	.000	.001	.000	.000	.000
	LDF	Mean	.013	.008	.009	.005	.007	.003
		Std	.012	.007	.011	.005	.008	.004
	LR	Mean	.015	.009	.012	.005	.009	.004
		Std	.014	.008	.013	.006	.011	.005
	KM	Mean	.497	.499	.494	.504	.499	.502
		Std	.051	.037	.049	.035	.046	.038
0.25 : 0.75	LPM	Mean	.088	.074	.048	.034	.014	.009
		Std	.033	.025	.026	.015	.014	.007
	LDF	Mean	.104	.091	.071	.055	.039	.030
		Std	.035	.027	.032	.019	.024	.014
	LR	Mean	.106	.092	.072	.056	.041	.032
		Std	.036	.027	.032	.019	.025	.015
	KM	Mean	.497	.499	.489	.494	.508	.498
		Std	.054	.036	.056	.038	.048	.039
0.50 : 0.50	LPM	Mean	.525	.516	.346	.343	.209	.194
		Std	.069	.050	.054	.041	.050	.031
	LDF	Mean	.522	.511	.346	.343	.211	.197
		Std	.068	.049	.054	.041	.050	.031
	LR	Mean	.520	.511	.347	.343	.212	.198
		Std	.067	.050	.054	.041	.050	.031
	KM	Mean	.500	.494	.491	.493	.492	.494
		Std	.057	.048	.063	.043	.057	.046

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.007	.006	.003	.003	.001	.001
		Std	.007	.004	.004	.003	.002	.001
	LDF	Mean	.027	.026	.023	.021	.019	.017
		Std	.014	.009	.012	.009	.012	.007
	LR	Mean	.029	.025	.029	.023	.027	.020
		Std	.014	.009	.014	.009	.016	.009
	KM	Mean	.511	.510	.511	.506	.505	.505
		Std	.049	.034	.050	.033	.049	.035
0.25 : 0.75	LPM	Mean	.076	.072	.059	.055	.041	.035
		Std	.025	.018	.021	.015	.017	.013
	LDF	Mean	.080	.077	.069	.064	.056	.050
		Std	.025	.019	.021	.017	.019	.015
	LR	Mean	.059	.053	.054	.048	.050	.040
		Std	.022	.015	.019	.015	.019	.014
	KM	Mean	.491	.504	.504	.514	.504	.513
		Std	.069	.054	.064	.044	.063	.042
0.50 : 0.50	LPM	Mean	.197	.189	.165	.159	.134	.127
		Std	.045	.029	.041	.027	.037	.026
	LDF	Mean	.196	.188	.165	.159	.134	.127
		Std	.045	.029	.041	.027	.037	.027
	LR	Mean	.139	.131	.116	.109	.096	.090
		Std	.041	.025	.036	.023	.032	.022
	KM	Mean	.456	.463	.480	.475	.489	.490
		Std	.085	.070	.084	.071	.078	.064

Covariance matrices = 1:4 and Group separation = 2.2								
			Sample representativeness					
Population proportions	Method		Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.003	.002	.001	.000	.000	.000
		Std	.005	.003	.002	.001	.001	.000
	LDF	Mean	.033	.029	.024	.020	.017	.014
		Std	.017	.012	.014	.010	.014	.008
	LR	Mean	.049	.042	.039	.033	.031	.024
		Std	.022	.016	.019	.013	.019	.012
	KM	Mean	.513	.506	.509	.505	.508	.503
		Std	.051	.036	.051	.036	.048	.037
0.25 : 0.75	LPM	Mean	.121	.120	.086	.080	.043	.036
		Std	.032	.022	.028	.022	.021	.013
	LDF	Mean	.132	.131	.107	.102	.073	.065
		Std	.034	.022	.030	.024	.026	.018
	LR	Mean	.130	.128	.113	.106	.086	.077
		Std	.032	.021	.029	.025	.028	.018
	KM	Mean	.517	.509	.514	.512	.522	.517
		Std	.055	.038	.049	.036	.052	.036
0.50 : 0.50	LPM	Mean	.363	.353	.296	.287	.215	.208
		Std	.055	.040	.051	.035	.045	.032
	LDF	Mean	.362	.352	.296	.287	.216	.209
		Std	.055	.040	.051	.035	.045	.033
	LR	Mean	.324	.314	.262	.257	.199	.093
		Std	.054	.038	.049	.033	.042	.031
	KM	Mean	.516	.514	.527	.522	.526	.517
		Std	.059	.041	.060	.045	.061	.043

Covariance matrices = 1:4 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.002	.000	.000	.000	.000	.000
	LDF	Mean	.013	.008	.007	.005	.004	.002
		Std	.012	.007	.008	.006	.006	.003
	LR	Mean	.021	.012	.013	.009	.009	.004
		Std	.017	.011	.012	.008	.011	.005
	KM	Mean	.507	.509	.506	.506	.510	.505
		Std	.047	.035	.050	.035	.051	.036
0.25 : 0.75	LPM	Mean	.120	.106	.065	.050	.020	.013
		Std	.038	.028	030	.021	.016	.010
	LDF	Mean	.138	.124	.093	.075	.048	.038
		Std	.040	.029	.036	.025	.026	.017
	LR	Mean	.149	.134	.107	.087	.062	.048
		Std	.041	.030	.036	.028	.032	.021
	KM	Mean	.519	.515	.519	.511	.522	.511
		Std	.052	.036	.054	.039	.046	.036
0.50 : 0.50	LPM	Mean	.505	.503	.391	.375	.259	.237
		Std	.052	.039	.052	.042	.053	.040
	LDF	Mean	.503	.501	.391	.375	.261	.240
		Std	.051	.039	.052	.042	.053	.040
	LR	Mean	.485	.485	.379	.364	.261	.240
		Std	.051	.039	.052	.041	.051	.039
	KM	Mean	.524	.518	.517	.518	.528	.515
		Std	.057	.041	.056	.044	.061	.042

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.002	.001	.001	.000	.000	.000
		Std	.004	.002	.002	.001	.001	.001
	LDF	Mean	.011	.009	.010	.008	.008	.007
		Std	.009	.006	.008	.005	.008	.005
	LR	Mean	.018	.012	.016	.010	.014	.009
		Std	.013	.007	.013	.007	.012	.006
	KM	Mean	.437	.468	.453	.469	.459	.476
		Std	.096	.046	.071	.052	.075	.039
0.25 : 0.75	LPM	Mean	.016	.014	.012	.010	.007	.006
		Std	.012	.008	.011	.006	.007	.005
	LDF	Mean	.017	.016	.015	.013	.012	.011
		Std	.013	.008	.012	.007	.010	.007
	LR	Mean	.041	.037	.035	.030	.027	.022
		Std	.020	.013	.018	.012	.015	.011
	KM	Mean	.268	.369	.304	.359	.325	.402
		Std	.142	.106	.150	.126	.148	.116
0.50 : 0.50	LPM	Mean	.040	.039	.023	.022	.014	.013
		Std	.020	.016	.016	.012	.013	.008
	LDF	Mean	.039	.039	.023	.022	.014	.013
		Std	.020	.016	.016	.012	.013	.008
	LR	Mean	.093	.089	.066	.065	.048	.044
		Std	.034	.023	.027	.020	.024	.015
	KM	Mean	.186	.236	.169	.190	.145	.163
		Std	.129	.116	.120	.120	.121	.117

Covariance matrices = 4:1 and Group separation = 2.2

Population Proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.002	.001	.001	.000	.000	.000
		Std	.003	.001	.002	.001	.001	.000
	LDF	Mean	.016	.010	.012	.009	.009	.008
		Std	.012	.006	.010	.007	.009	.006
	LR	Mean	.015	.007	.013	.006	.010	.005
		Std	.012	.006	.012	.006	.011	.005
	KM	Mean	.466	.476	.469	.474	.468	.479
		Std	.061	.039	.058	.044	.057	.043
0.25 : 0.75	LPM	Mean	.039	.031	.024	.019	.012	.009
		Std	.019	.013	.016	.010	.010	.006
	LDF	Mean	.046	.038	.034	.029	.025	.022
		Std	.021	.014	.019	.012	.015	.010
	LR	Mean	.060	.048	.043	.034	.029	.023
		Std	.023	.016	.023	.014	.017	.012
	KM	Mean	.431	.454	.435	.457	.447	.463
		Std	.087	.069	.073	.056	.074	.054
0.50 : 0.50	LPM	Mean	.242	.244	.121	.121	.065	.062
		Std	.059	.043	.038	.025	.028	.019
	LDF	Mean	.239	.241	.121	.121	.066	.063
		Std	.058	.042	.038	.025	.028	.019
	LR	Mean	.279	.278	.165	.165	.101	.093
		Std	.059	.044	.044	.028	.033	.024
	KM	Mean	.406	.413	.397	.422	.412	.423
		Std	.114	.089	.113	.085	.104	.088

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.000	.000	.000	.000	.000	.000
		Std	.002	.000	.001	.000	.000	.000
	LDF	Mean	.010	.005	008	.004	.008	.003
		Std	.011	.005	.009	.005	.008	.005
	LR	Mean	.008	.003	.007	.002	.008	.002
		Std	.011	.005	.009	.004	.011	.003
	KM	Mean	.469	.479	.477	.480	.470	.488
		Std	.057	.039	.057	.043	.057	.040
0.25 : 0.75	LPM	Mean	.037	.023	.018	.009	.007	.003
		Std	.023	.013	.016	.007	.007	.005
	LDF	Mean	.046	.032	.031	.019	.020	.013
		Std	.024	.014	.022	.012	.016	.010
	LR	Mean	.049	.033	.032	.018	.018	.010
		Std	.027	.016	.022	.012	.014	.009
	KM	Mean	.448	.466	.457	.473	.449	.476
		Std	.075	.057	.072	.056	.067	.049
0.50 : 0.50	LPM	Mean	.525	.537	.248	.245	.106	.093
		Std	.088	.076	.057	.038	.035	.025
	LDF	Mean	.521	.531	.248	.245	.107	.096
		Std	.087	.075	.057	.038	.036	.025
	LR	Mean	.520	.529	.270	.266	.125	.111
		Std	.083	.071	.060	.038	.039	.028
	KM	Mean	.444	.456	.434	.460	.435	.457
		Std	.096	.068	.095	.069	.088	.060

Table G2-3

ANOVA Results, Eta-squared, and Partial-Omega squared of Group 2 Error Rate for Data Pattern I on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	398.99	199.50	47059.40	<.0001	.1252	.4207
Equality of covariance (COV)	2	183.22	91.61	21609.60	<.0001	.0575	.2401
Group separation (GS)	2	295.45	147.72	34846.50	<.0001	.0927	.3497
Sample representativeness (SR)	2	58.11	29.05	6853.52	<.0001	.0182	.0956
Sample size (SS)	1	0.0043	0.0043	1.02	.3125	<.0001	<.0001
PP*COV	4	44.43	11.11	2620.28	<.0001	.0139	.0748
PP*GS	4	151.09	37.77	8909.89	<.0001	.0474	.2157
PP*SR	4	77.90	19.48	4594.08	<.0001	.0244	.1242
PP*SS	2	0.0197	0.0099	2.33	.0977	<.0001	<.0001
COV*GS	4	3.68	0.92	216.78	<.0001	.0012	.0066
COV*SR	4	0.53	0.13	31.04	<.0001	.0002	.0009
COV*SS	2	0.018	0.0088	2.09	.1241	<.0001	<.0001
GS*SR	4	27.50	6.88	1621.77	<.0001	.0086	.0476
GS*SS	2	0.0005	0.0003	0.06	.9390	<.0001	<.0001
SR*SS	2	0.014	0.007	1.66	.1910	<.0001	<.0001
Error (between)	32358	137.17	0.0042				
Within-subject effects:							
Method (4M)	3	850.91	283.64	324376.0	<.0001	.2669	.8825
4M*PP	6	530.42	88.40	101101.0	<.0001	.1664	.8240
4M*COV	6	105.41	17.57	20092.10	<.0001	.0331	.4819
4M*GS	6	81.75	13.63	15582.80	<.0001	.0256	.4191
4M*SR	6	40.55	6.76	7729.08	<.0001	.0127	.2635
4M*SS	3	0.0493	0.0164	18.80	<.0001	<.0001	.0004
4M*PP*COV	12	10.81	0.90	1029.97	<.0001	.0034	.0870
4M*PP*GS	12	72.15	6.01	6876.40	<.0001	.0226	.3890
4M*PP*SR	12	22.38	1.87	2133.14	<.0001	.0070	.1649
4M*PP*SS	6	0.040	0.0066	7.57	<.0001	<.0001	.0003
4M*COV*GS	12	1.84	0.15	175.71	<.0001	.0006	.0159
4M*COV*SR	12	0.46	0.039	44.26	<.0001	.0001	.0040
4M*COV*SS	6	0.0071	0.0012	1.36	.2274	<.0001	<.0001
4M*GS*SR	12	8.10	0.67	771.85	<.0001	.0025	.0666
4M*GS*SS	6	0.0028	0.0005	0.54	.7812	<.0001	<.0001
4M*SR*SS	6	0.0076	0.0013	1.44	.1940	<.0001	<.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	84.88	0.0009				

Note. Eta square (η^2) is defined as: $\eta^2 = \dfrac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \dfrac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table G2-4

ANOVA Results, Eta-squared, and Partial-Omega squared of Group 2 Error Rate for Data Pattern II on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	510.26	255.13	60414.90	<.0001	.0976	.4825
Equality of covariance (COV)	2	97.39	48.70	11531.40	<.0001	.0186	.1511
Group separation (GS)	2	140.77	70.38	16667.00	<.0001	.0269	.2046
Sample representativeness (SR)	2	66.17	33.08	7834.55	<.0001	.0127	.1079
Sample size (SS)	1	0.11	0.11	25.19	<.0001	<.0001	.0002
PP*COV	4	38.92	9.73	2303.83	<.0001	.0074	.0664
PP*GS	4	209.60	52.40	12408.50	<.0001	.0401	.2769
PP*SR	4	69.40	17.35	4108.51	<.0001	.0133	.1125
PP*SS	2	0.0070	0.0035	0.82	.4385	<.0001	<.0001
COV*GS	4	4.22	1.06	249.88	<.0001	.0008	.0076
COV*SR	4	0.55	0.14	32.59	<.0001	.0001	.0010
COV*SS	2	0.47	0.23	55.25	<.0001	.0001	.0008
GS*SR	4	24.43	6.11	1446.17	<.0001	.0047	.0427
GS*SS	2	0.2807	0.1404	33.24	<.0001	.0001	.0005
SR*SS	2	0.015	0.0077	1.82	.1619	<.0001	<.0001
Error (between)	32358	136.65	0.0042				
Within-subject effects:							
Method (4M)	3	3363.46	1121.15	723635.0	<.0001	.6435	.9437
4M*PP	6	305.18	50.86	32829.20	<.0001	.0584	.6031
4M*COV	6	9.45	1.57	1016.03	<.0001	.0018	.0449
4M*GS	6	2.92	0.49	314.06	<.0001	.0006	.0143
4M*SR	6	25.70	4.28	2764.23	<.0001	.0049	.1134
4M*SS	3	1.21	0.40	260.27	<.0001	.0002	.0060
4M*PP*COV	12	3.98	0.33	213.86	<.0001	.0008	.0193
4M*PP*GS	12	21.66	1.80	1164.87	<.0001	.0041	.0973
4M*PP*SR	12	21.46	1.79	1154.08	<.0001	.0041	.0965
4M*PP*SS	6	0.21	0.035	22.78	<.0001	<.0001	.0010
4M*COV*GS	12	14.22	1.19	764.87	<.0001	.0027	.0661
4M*COV*SR	12	0.26	0.022	13.88	<.0001	<.0001	.0012
4M*COV*SS	6	0.66	0.11	70.73	<.0001	.0001	.0032
4M*GS*SR	12	6.70	0.56	360.24	<.0001	.0013	.0322
4M*GS*SS	6	0.24	0.039	25.29	<.0001	<.0001	.0011
4M*SR*SS	6	0.0067	0.0011	0.72	.6328	<.0001	<.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	150.40	0.0015				

Note. Eta square (η^2) is defined as: $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table G2-5

Means and Standard Deviations of Group 2 Error Rate for Method by Population Proportion Interaction

Population proportions	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
0.10:0.90	.00086	.0026	.011	.011	.013	.014	.360	.116
0.25:0.75	.038	.034	.051	.037	.053	.037	.263	.147
0.50:0.50	.218	.153	.218	.152	.218	.145	.211	.154
				Data Pattern II				
0.10:0.90	.0011	.0029	.014	.012	.017	.016	.491	.051
0.25:0.75	.044	.037	.057	.040	.061	.040	.471	.088
0.50:0.50	.221	.151	.221	.150	.222	.142	.439	.129

Table G2-6

Means and Standard Deviations of Group 2 Error Rate for Method by Equality of Covariance Matrices Interaction

Equality of covariance matrices	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
1:4	.109	.136	.119	.131	.112	.123	.374	.109
Equal	.091	.129	.099	.125	.100	.125	.296	.122
4:1	.056	.121	.060	.118	.071	.121	.164	.145
				Data Pattern II				
1:4	.115	.138	.126	.132	.119	.123	.508	.053
Equal	.094	.130	.104	.125	.106	.124	.482	.065
4:1	.057	.116	.062	.113	.075	.117	.411	.126

Table G2-7

Means and Standard Deviations of Group 2 Error Rate for Method by Group Separation Interaction

Group separation	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	Data Pattern I							
6.7	.043	.055	.050	.052	.050	.041	.158	.108
2.2	.090	.112	.099	.107	.103	.103	.296	.137
0.7	.124	.181	.130	.177	.132	.174	.379	.121
	Data Pattern II							
6.7	.045	.057	.052	.054	.053	.042	.426	.133
2.2	.093	.113	.104	.107	.108	.102	.484	.068
0.7	.128	.178	.136	.173	.138	.171	.491	.059

Table G2-8

Means and Standard Deviations of Group 2 Error Rate for Method by Sample Representativeness Interaction

Sample representativeness	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
20% over	.127	.174	.133	.169	.134	.165	.267	.152
Equal	.082	.114	.089	.110	.091	.108	.277	.153
20% under	.048	.072	.057	.070	.059	.068	.290	.152
				Data Pattern II				
20% over	.129	.171	.136	.165	.137	.161	.465	.096
Equal	.085	.116	.094	.112	.097	.108	.466	.098
20% under	.051	.076	.061	.073	.065	.070	.469	.097

Table G2-9

Means and Standard Deviations of Group 2 Error Rate for Method by Population Proportions by Equality of Covariance Matrices Interaction

Population proportions	Equality of covariance matrices	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
					Data Pattern I				
0.10:0.90	1:4	.0014	.0034	.015	.013	.019	.016	.407	.089
	Equal	.0009	.0026	.012	.011	.013	.012	.379	.088
	4:1	.0003	.0012	.006	.007	.005	.007	.294	.133
0.25:0.75	1:4	.059	.038	.076	.037	.075	.040	.361	.116
	Equal	.042	.029	.057	.030	.058	.031	.287	.114
	4:1	.013	.013	.020	.015	.026	.019	.140	.114
0.50:0.50	1:4	.268	.122	.268	.121	.241	.129	.354	.111
	Equal	.230	.140	.229	.139	.230	.139	.221	.106
	4:1	.156	.170	.155	.168	.183	.158	.058	.060
					Data Pattern II				
0.10:0.90	1:4	.0015	.0035	.017	.014	.024	.018	.507	.043
	Equal	.0013	.0031	.015	.012	.018	.014	.496	.044
	4:1	.0005	.0017	.009	.008	.009	.010	.470	.057
0.25:0.75	1:4	.067	.040	.084	.039	.085	.043	.512	.050
	Equal	.047	.031	.063	.032	.065	.033	.486	.055
	4:1	.017	.016	.024	.018	.033	.021	.416	.112
0.50:0.50	1:4	.276	.121	.276	.121	.247	.129	.505	.065
	Equal	.233	.140	.233	.139	.234	.138	.464	.086
	4:1	.153	.162	.153	.160	.184	.150	.347	.157

Table G2-10

Means and Standard Deviations of Group 2 Error Rate for Method by Population Proportions by Group Separation Interaction

		Method							
Group separation	Population proportions	LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		Data Pattern I							
6.7	0.10:0.90	.002	.004	.015	.011	.016	.011	.240	.096
	0.25:0.75	.033	.023	.040	.026	.039	.019	.128	.086
	0.50:0.50	.095	.063	.095	.063	.093	.039	.108	.092
2.2	0.10:0.90	.0006	.002	.014	.012	.017	.017	.390	.069
	0.25:0.75	.049	.038	.065	.039	.068	.039	.279	.122
	0.50:0.50	.220	.099	.220	.098	.223	.085	.220	.148
0.7	0.10:0.90	.00003	.0004	.004	.006	.005	.008	.449	.056
	0.25:0.75	.032	.036	.048	.040	.051	.044	.383	.098
	0.50:0.50	.339	.163	.338	.161	.339	.154	.307	.145
		Data Pattern II							
6.7	0.10:0.90	.002	.004	.016	.011	.020	.013	.486	.056
	0.25:0.75	.035	.026	.042	.028	.044	.020	.437	.119
	0.50:0.50	.097	.066	.097	.066	.096	.040	.355	.164
2.2	0.10:0.90	.0009	.002	.018	.013	.023	.019	.492	.049
	0.25:0.75	.054	.039	.070	.041	.075	.040	.485	.062
	0.50:0.50	.2232	.099	.2229	.099	.226	.084	.474	.086
0.7	0.10:0.90	.00009	.0008	.007	.008	.008	.011	.494	.047
	0.25:0.75	.041	.041	.059	.044	.064	.047	.492	.057
	0.50:0.50	.342	.155	.342	.153	.342	.146	.487	.070

Table G2-11

Means and Standard Deviations of Group 2 Error Rate for Method by Population Proportions by Sample Representativeness Interaction

Sample representativeness	Population proportions	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I					
20% over	0.10:0.90	.002	.004	.014	.013	.016	.016	.346	.123
	0.25:0.75	.061	.040	.070	.043	.072	.042	.253	.147
	0.50:0.50	.318	.181	.315	.179	.313	.172	.202	.149
Equal	0.10:0.90	.0006	.002	.011	.011	.013	.013	.358	.117
	0.25:0.75	.036	.027	.049	.032	.051	.033	.262	.147
	0.50:0.50	.208	.116	.208	.116	.209	.108	.212	.155
20%under	0.10:0.90	.0001	.0008	.008	.009	.009	.011	.376	.105
	0.25:0.75	.017	.017	.033	.024	.035	.026	.275	.146
	0.50:0.50	.128	.076	.129	.076	.132	.069	.220	.157
				Data Pattern II					
20% over	0.10:0.90	.002	.004	.017	.014	.021	.018	.490	.054
	0.25:0.75	.068	.042	.077	.046	.080	.044	.466	.092
	0.50:0.50	.316	.176	.314	.174	.312	.167	.439	.122
Equal	0.10:0.90	.0008	.002	.013	.012	.017	.015	.491	.050
	0.25:0.75	.042	.030	.057	.036	.061	.036	.470	.088
	0.50:0.50	.212	.119	.212	.119	.214	.110	.438	.130
20%under	0.10:0.90	.0002	.0009	.010	.010	.014	.014	.492	.049
	0.25:0.75	.020	.018	.038	.025	.042	.027	.478	.082
	0.50:0.50	.134	.081	.135	.081	.138	.073	.439	.133

Table G2-12

Means and Standard Deviations of Group 2 Error Rate for Method by Equality of Covariance Matrices by Group Separation Interaction

Group separation	Equality of covariance matrices	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I					
6.7	1:4	.070	.069	.079	.062	.058	.043	.245	.059
	Equal	.048	.048	.056	.043	.057	.044	.175	.087
	4:1	.012	.015	.015	.014	.034	.031	.055	.077
2.2	1:4	.119	.129	.132	.120	.128	.104	.406	.049
	Equal	.097	.112	.108	.106	.109	.105	.317	.085
	4:1	.054	.081	.059	.077	.071	.091	.165	.132
0.7	1:4	.139	.179	.148	.174	.150	.167	.472	.048
	Equal	.128	.179	.135	.175	.136	.174	.396	.067
	4:1	.103	.182	.106	.178	.109	.179	.271	.128
				Data Pattern II					
6.7	1:4	.074	.072	.083	.065	.063	.044	.496	.062
	Equal	.048	.049	.056	.043	.059	.043	.456	.084
	4:1	.012	.015	.016	.014	.037	.031	.326	.162
2.2	1:4	.123	.130	.137	.121	.134	.102	.514	.047
	Equal	.100	.113	.112	.105	.114	.104	.493	.049
	4:1	.055	.080	.062	.075	.076	.090	.444	.080
0.7	1:4	.147	.181	.157	.175	.159	.167	.514	.047
	Equal	.133	.179	.142	.173	.143	.172	.497	.047
	4:1	.103	.173	.108	.169	.112	.171	.462	.067

Table G2-13

Means and Standard Deviations of Group 2 Error Rate for Method by Group Separation by Sample Representativeness Interaction

Group separation	Sample representativeness	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I					
6.7	20% over	.055	.063	.061	.059	.061	.047	.146	.100
	Equal	.043	.053	.049	.050	.049	.039	.156	.106
	20% under	.032	.045	.040	.042	.039	.034	.173	.116
2.2	20% over	.128	.138	.137	.131	.139	.127	.285	.138
	Equal	.087	.103	.097	.098	.100	.094	.296	.138
	20% under	.055	.074	.065	.071	.069	.068	.307	.136
0.7	20% over	.197	.240	.202	.235	.201	.231	.369	.123
	Equal	.115	.151	.122	.147	.125	.146	.379	.121
	20% under	.059	.088	.065	.086	.068	.087	.390	.117
				Data Pattern II					
6.7	20% over	.058	.066	.064	.062	.0646	.049	.422	.127
	Equal	.044	.055	.051	.052	.053	.039	.425	.134
	20% under	.032	.045	.040	.042	.042	.032	.431	.136
2.2	20% over	.131	.138	.140	.130	.143	.125	.482	.070
	Equal	.090	.105	.101	.099	.106	.093	.483	.068
	20% under	.057	.076	.069	.071	.075	.068	.486	.066
0.7	20% over	.198	.233	.204	.227	.204	.223	.491	.060
	Equal	.121	.152	.130	.148	.133	.147	.490	.059
	20% under	.065	.095	.074	.092	.077	.092	.492	.058

Table T-1

Means and Standard Deviations of Total Error Rates of 200 Replications for Data Pattern I

Covariance matrices = Equal and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.057	.055	.066	.067	.079	.081
		Std	.017	.012	.019	.013	.020	.014
	LDF	Mean	.050	.048	.049	.049	.050	.049
		Std	.015	.011	.015	.011	.017	.010
	LR	Mean	.051	.048	.051	.049	.051	.049
		Std	.016	.011	.015	.010	.017	.011
	KM	Mean	.235	.233	.246	.253	.271	.286
		Std	.054	.038	.051	.039	.053	.040
0.25 : 0.75	LPM	Mean	.082	.082	.083	.081	.090	.091
		Std	.018	.013	.019	.014	.021	.015
	LDF	Mean	.083	.082	.084	.080	.085	.084
		Std	.018	.013	.019	.015	.020	.014
	LR	Mean	.083	.083	.085	.081	.086	.085
		Std	.020	.013	.019	.015	.020	.014
	KM	Mean	.119	.121	.127	.124	.136	.135
		Std	.030	.021	.030	.021	.034	.022
0.50 : 0.50	LPM	Mean	.102	.100	.101	.100	.101	.098
		Std	.019	.015	.022	.013	.021	.016
	LDF	Mean	.101	.100	.101	.100	.101	.098
		Std	.019	.015	.022	.013	.021	.016
	LR	Mean	.103	.100	.103	.101	.102	.099
		Std	.020	.015	.023	.014	.021	.015
	KM	Mean	.098	.097	.100	.100	.100	.099
		Std	.020	.014	.022	.014	.022	.016

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.095	.095	.097	.099	.101	.100
		Std	.022	.015	.022	.015	.019	.014
	LDF	Mean	.092	.090	.091	.091	.092	.089
		Std	.021	.014	.021	.014	.020	.014
	LR	Mean	.092	.090	.092	.091	.093	.089
		Std	.021	.014	.022	.014	.020	.013
	KM	Mean	.360	.376	.375	.377	.385	.384
		Std	.047	.036	.049	.037	.044	.035
0.25 : 0.75	LPM	Mean	.179	.178	.179	.176	.192	.191
		Std	.031	.018	.030	.020	.032	.020
	LDF	Mean	.180	.179	.177	.175	.181	.181
		Std	.031	.018	.030	.020	.029	.020
	LR	Mean	.181	.179	.178	.175	.181	.180
		Std	.031	.018	.029	.019	.028	.020
	KM	Mean	.268	.269	.272	.276	.289	.284
		Std	.039	.025	.037	.029	.040	.029
0.50 : 0.50	LPM	Mean	.239	.240	.230	.230	.238	.237
		Std	.032	.022	.029	.022	.032	.021
	LDF	Mean	.239	.240	.230	.230	.237	.237
		Std	.032	.022	.029	.022	.032	.021
	LR	Mean	.239	.239	.230	.230	.237	.236
		Std	.032	.022	.029	.022	.032	.021
	KM	Mean	.231	.232	.229	.230	.228	.229
		Std	.032	.022	.030	.022	.030	.021

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.102	.099	.100	.101	.100	.101
		Std	.022	.016	.020	.016	.021	.014
	LDF	Mean	.103	.100	.101	.102	.101	.100
		Std	.022	.016	.021	.016	.020	.014
	LR	Mean	.104	.100	.101	.102	.102	.100
		Std	.022	.017	.021	.017	.020	.014
	KM	Mean	.422	.425	.432	.433	.434	.436
		Std	.048	.035	.046	.030	.045	.032
0.25 : 0.75	LPM	Mean	.237	.237	.240	.239	.234	.242
		Std	.030	.022	.028	.022	.032	.020
	LDF	Mean	.240	.239	.243	.236	.239	.237
		Std	.031	.022	.028	.021	.031	.020
	LR	Mean	.241	.240	.243	.237	.239	.236
		Std	.031	.022	.028	.020	.032	.020
	KM	Mean	.365	.365	.375	.363	.372	.377
		Std	.037	.026	.038	.027	.041	.027
0.50 : 0.50	LPM	Mean	.359	.355	.337	.335	.354	.353
		Std	.039	.025	.034	.026	.037	.025
	LDF	Mean	.358	.354	.337	.335	.353	.353
		Std	.039	.024	.034	.026	.036	.025
	LR	Mean	.358	.354	.337	.335	.354	.352
		Std	.039	.024	.034	.027	.037	.025
	KM	Mean	.336	.337	.334	.334	.335	.334
		Std	.033	.024	.032	.024	.033	.025

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.057	.061	.080	.079	.095	.095
		Std	.018	.015	.022	.016	.022	.015
	LDF	Mean	.037	.037	.039	.038	.043	.039
		Std	.014	.010	.015	.010	.015	.010
	LR	Mean	.039	.038	.041	.039	.044	.039
		Std	.013	.010	.016	.010	.015	.010
	KM	Mean	.248	.256	.266	.272	.292	.296
		Std	.049	.032	.049	.036	.051	.037
0.25 : 0.75	LPM	Mean	.066	.062	.060	.059	.068	.065
		Std	.017	.013	.017	.011	.019	.013
	LDF	Mean	.067	.064	.062	.061	.061	.059
		Std	.017	.013	.017	.012	.017	.012
	LR	Mean	.062	.058	.062	.060	.066	.063
		Std	.016	.012	.017	.011	.018	.013
	KM	Mean	.153	.152	.159	.158	.172	.173
		Std	.032	.022	.034	.024	.031	.025
0.50 : 0.50	LPM	Mean	.100	.095	.090	.086	.085	.081
		Std	.022	.015	.022	.014	.020	.013
	LDF	Mean	.100	.095	.090	.086	.084	.081
		Std	.022	.015	.022	.014	.020	.014
	LR	Mean	.087	.080	.086	.081	.092	.087
		Std	.020	.014	.021	.015	.022	.014
	KM	Mean	.111	.109	.112	.111	.114	.116
		Std	.023	.017	.023	.017	.025	.019

Covariance matrices = 1:4 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.101	.102	.100	.100	.102	.101
		Std	.022	.014	.022	.015	.022	.015
	LDF	Mean	.109	.110	.108	.106	.108	.106
		Std	.022	.015	.023	.015	.021	.015
	LR	Mean	.112	.112	.112	.108	.110	.108
		Std	.023	.016	.024	.016	.022	.014
	KM	Mean	.392	.384	.394	.392	.390	.397
		Std	.046	.030	.046	.034	.044	.033
0.25 : 0.75	LPM	Mean	.183	.184	.205	.204	.229	.236
		Std	.030	.019	.031	.020	.031	.023
	LDF	Mean	.181	.181	.194	.192	.208	.212
		Std	.028	.019	.030	.019	.028	.023
	LR	Mean	.182	.182	.191	.189	.202	.205
		Std	.029	.019	.029	.018	.027	.023
	KM	Mean	.297	.294	.303	.302	.304	.306
		Std	.038	.030	.039	.027	.039	.024
0.50 : 0.50	LPM	Mean	.209	.206	.206	.201	.225	.224
		Std	.028	.021	.030	.020	.033	.023
	LDF	Mean	.209	.206	.206	.201	.224	.223
		Std	.028	.021	.030	.020	.032	.023
	LR	Mean	.206	.202	.212	.206	.235	.233
		Std	.028	.019	.030	.020	.033	.023
	KM	Mean	.214	.212	.215	.214	.214	.216
		Std	.027	.023	.030	.021	.029	.024

Covariance matrices = 1:4 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.099	.099	.099	.100	.100	.100
		Std	.021	.015	.022	.016	.020	.016
	LDF	Mean	.107	.105	.103	.103	.102	.101
		Std	.024	.016	.023	.016	.021	.016
	LR	Mean	.112	.108	.106	.105	.104	.102
		Std	.026	.017	.024	.017	.022	.016
	KM	Mean	.441	.438	.438	.444	.441	.447
		Std	.042	.031	.042	.031	.044	.029
0.25 : 0.75	LPM	Mean	.287	.285	.276	.273	.258	.257
		Std	.030	.022	.033	.021	.030	.022
	LDF	Mean	.290	.287	.282	.281	.269	.271
		Std	.030	.022	.034	.021	.031	.024
	LR	Mean	.291	.288	.285	.283	.273	.275
		Std	.030	.022	.034	.022	.031	.024
	KM	Mean	.374	.378	.376	.374	.376	.377
		Std	.037	.025	.037	.027	.036	.028
0.50 : 0.50	LPM	Mean	.298	.297	.309	.311	.392	.393
		Std	.033	.023	.033	.026	.042	.033
	LDF	Mean	.298	.297	.309	.311	.390	.391
		Std	.033	.023	.033	.026	.042	.032
	LR	Mean	.297	.296	.313	.314	.389	.390
		Std	.033	.023	.034	.026	.041	.031
	KM	Mean	.294	.295	.295	.294	.295	.295
		Std	.033	.023	.030	.024	.031	.024

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.055	.054	.061	.059	.068	.068
		Std	.017	.012	.017	.013	.020	.014
	LDF	Mean	.049	.047	.048	.047	.049	.049
		Std	.015	.010	.016	.011	.017	.012
	LR	Mean	.050	.048	.050	.048	.050	.051
		Std	.016	.011	.016	.011	.017	.012
	KM	Mean	.120	.117	.140	.135	.171	.187
		Std	.061	.042	.070	.060	.082	.070
0.25 : 0.75	LPM	Mean	.081	.079	.084	.081	.089	.088
		Std	.020	.014	.020	.012	.021	.014
	LDF	Mean	.081	.078	.081	.078	.082	.081
		Std	.019	.014	.019	.012	.020	.013
	LR	Mean	.083	.079	.079	.077	.079	.077
		Std	.020	.013	.019	.013	.020	.013
	KM	Mean	.079	.077	.077	.076	.077	.076
		Std	.019	.013	.019	.013	.019	.013
0.50 : 0.50	LPM	Mean	.080	.083	.091	.086	.100	.097
		Std	.020	.014	.021	.014	.020	.015
	LDF	Mean	.081	.083	.091	.086	.100	.097
		Std	.020	.015	.021	.014	.020	.015
	LR	Mean	.088	.089	.087	.082	.086	.083
		Std	.021	.014	.021	.014	.020	.014
	KM	Mean	.116	.117	.114	.110	.114	.109
		Std	.024	.018	.023	.017	.023	.017

Covariance matrices = 4:1 and Group separation = 2.2

Population Proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.086	.085	.088	.088	.093	.093
		Std	.022	.014	.022	.016	.020	.015
	LDF	Mean	.075	.073	.074	.073	.076	.073
		Std	.019	.013	.020	.013	.019	.014
	LR	Mean	.076	.075	.076	.075	.078	.075
		Std	.020	.013	.020	.013	.019	.014
	KM	Mean	.300	.313	.320	.327	.340	.345
		Std	.067	.051	.067	.049	.066	.045
0.25 : 0.75	LPM	Mean	.143	.147	.151	.149	.158	.156
		Std	.025	.019	.027	.019	.027	.019
	LDF	Mean	.143	.147	.148	.147	.149	.146
		Std	.026	.019	.026	.018	.026	.019
	LR	Mean	.146	.149	.148	.147	.150	.146
		Std	.026	.019	.026	.018	.027	.018
	KM	Mean	.171	.172	.183	.181	.191	.188
		Std	.036	.024	.034	.028	.041	.029
0.50 : 0.50	LPM	Mean	.229	.226	.203	.202	.204	.204
		Std	.031	.020	.027	.021	.028	.022
	LDF	Mean	.228	.225	.203	.202	.204	.204
		Std	.032	.021	.027	.021	.028	.022
	LR	Mean	.237	.235	.207	.208	.200	.200
		Std	.032	.021	.028	.020	.029	.022
	KM	Mean	.217	.216	.213	.214	.207	.211
		Std	.032	.022	.030	.020	.030	.022

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.095	.095	.098	.099	.099	.100
		Std	.021	.014	.023	.015	.020	.014
	LDF	Mean	.088	.086	.089	.089	.090	.091
		Std	.021	.015	.022	.015	.020	.015
	LR	Mean	.089	.087	.091	.091	.091	.093
		Std	.022	.015	.023	.015	.020	.014
	KM	Mean	.397	.400	.404	.407	.408	.419
		Std	.062	.043	.054	.043	.060	.040
0.25 : 0.75	LPM	Mean	.191	.187	.196	.197	.213	.209
		Std	.029	.021	.033	.021	.033	.025
	LDF	Mean	.190	.186	.191	.189	.200	.193
		Std	.028	.020	.031	.020	.032	.022
	LR	Mean	.191	.186	.192	.190	.201	.196
		Std	.028	.020	.032	.020	.033	.023
	KM	Mean	.291	.290	.302	.298	.317	.321
		Std	.049	.035	.049	.034	.052	.036
0.50 : 0.50	LPM	Mean	.393	.400	.312	.309	.297	.298
		Std	.041	.027	.034	.024	.033	.024
	LDF	Mean	.391	.398	.312	.309	.297	.297
		Std	.041	.027	.034	.024	.034	.024
	LR	Mean	.390	.396	.317	.314	.296	.296
		Std	.041	.026	.034	.024	.034	.023
	KM	Mean	.289	.296	.294	.292	.292	.294
		Std	.030	.022	.030	.023	.035	.023

Table T-2

Means and Standard Deviations of Total Error Rates of 200 Replications for Data Pattern II

Covariance matrices = Equal and Group separation = 6.7								
			Sample representativeness					
Population proportions	Method		Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.058	.057	.067	.066	.081	.080
		Std	.018	.012	.020	.013	.020	.015
	LDF	Mean	.052	.048	.052	.049	.052	.049
		Std	.016	.011	.015	.011	.015	.011
	LR	Mean	.057	.051	.058	.051	.056	.050
		Std	.016	.011	.017	.011	.017	.011
	KM	Mean	.483	.488	.483	.491	.495	.488
		Std	.055	.036	.045	.035	.046	.033
0.25 : 0.75	LPM	Mean	.086	.083	.085	.081	.094	.087
		Std	.021	.013	.018	.013	.022	.016
	LDF	Mean	.086	.084	.084	.080	.088	.081
		Std	.022	.013	.018	.013	.021	.015
	LR	Mean	.089	.085	.089	.082	.093	.083
		Std	.023	.014	.020	.014	.020	.015
	KM	Mean	.447	.462	.448	.474	.469	.468
		Std	.079	.059	.072	.047	.064	.052
0.50 : 0.50	LPM	Mean	.108	.104	.106	.100	.107	.104
		Std	.022	.016	.023	.016	.023	.016
	LDF	Mean	.108	.103	.106	.100	.107	.104
		Std	.022	.016	.023	.016	.023	.016
	LR	Mean	.110	.105	.108	.101	.109	.104
		Std	.023	.017	.024	.016	.024	.016
	KM	Mean	.396	.418	.396	.407	.406	.423
		Std	.098	.083	.112	.099	.101	.085

Covariance matrices = Equal and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.094	.094	.094	.096	.098	.100
		Std	.021	.015	.021	.015	.023	.016
	LDF	Mean	.096	.091	.091	.090	.093	.092
		Std	.024	.014	.022	.015	.022	.015
	LR	Mean	.099	.092	.094	.091	.094	.093
		Std	.024	.014	.021	.015	.023	.015
	KM	Mean	.501	.498	.498	.495	.494	.497
		Std	.045	.031	.044	.032	.049	.030
0.25 : 0.75	LPM	Mean	.181	.179	.184	.180	.192	.191
		Std	.028	.018	.028	.022	.027	.021
	LDF	Mean	.182	.181	.185	.179	.185	.182
		Std	.028	.018	.028	.021	.025	.021
	LR	Mean	.184	.182	.185	.180	.186	.181
		Std	.029	.018	.028	.021	.026	.021
	KM	Mean	.488	.490	.490	.489	.494	.492
		Std	.041	.031	.046	.034	.042	.033
0.50 : 0.50	LPM	Mean	.244	.241	.240	.234	.245	.243
		Std	.029	.023	.030	.020	.034	.023
	LDF	Mean	.244	.241	.240	.234	.245	.243
		Std	.029	.023	.030	.020	.034	.022
	LR	Mean	.245	.241	.240	.234	.246	.243
		Std	.029	.023	.031	.019	.033	.022
	KM	Mean	.489	.490	.491	.489	.488	.490
		Std	.043	.031	.041	.032	.046	.033

Covariance matrices = Equal and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.100	.098	.098	.100	.100	.100
		Std	.020	.015	.021	.015	.020	.014
	LDF	Mean	.105	.100	.102	.101	.103	.101
		Std	.021	.015	.021	.015	.020	.014
	LR	Mean	.106	.100	.104	.102	.104	.101
		Std	.022	.015	.021	.015	.021	.014
	KM	Mean	.498	.499	.494	.503	.497	.502
		Std	.046	.032	.045	.031	.042	.034
0.25 : 0.75	LPM	Mean	.245	.240	.243	.237	.243	.245
		Std	.029	.021	.031	.021	.031	.022
	LDF	Mean	.248	.243	.247	.237	.241	.240
		Std	.030	.022	.030	.021	.030	.022
	LR	Mean	.249	.244	.248	.237	.242	.240
		Std	.031	.022	.030	.021	.030	.022
	KM	Mean	.498	.498	.492	.495	.504	.498
		Std	.040	.027	.042	.028	.037	.029
0.50 : 0.50	LPM	Mean	.362	.353	.347	.342	.361	.354
		Std	.034	.026	.034	.026	.035	.024
	LDF	Mean	.362	.352	.347	.342	.361	.353
		Std	.034	.026	.034	.026	.035	.024
	LR	Mean	.361	.352	.347	.342	.361	.353
		Std	.034	.026	.035	.026	.036	.025
	KM	Mean	.499	.497	.493	.496	.492	.495
		Std	.037	.028	.036	.027	.038	.030

Covariance matrices = 1:4 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.056	.059	.077	.077	.095	.096
		Std	.018	.014	.020	.014	.021	.015
	LDF	Mean	.039	.038	.040	.038	.044	.041
		Std	.014	.010	.013	.010	.016	.010
	LR	Mean	.044	.041	.047	.040	.050	.043
		Std	.015	.010	.015	.010	.017	.011
	KM	Mean	.491	.496	.496	.492	.493	.492
		Std	.047	.033	.048	.032	.048	.032
0.25 : 0.75	LPM	Mean	.069	.065	.063	.061	.070	.066
		Std	.019	.014	.018	.012	.018	.014
	LDF	Mean	.071	.067	.065	.062	.065	.061
		Std	.019	.014	.017	.012	.017	.013
	LR	Mean	.069	.062	.068	.064	.073	.066
		Std	.019	.013	.019	.013	.019	.013
	KM	Mean	.415	.434	.432	.445	.441	.462
		Std	.071	.059	.065	.048	.070	.046
0.50 : 0.50	LPM	Mean	.105	.101	.094	.091	.086	.083
		Std	.024	.016	.021	.015	.020	.014
	LDF	Mean	.105	.100	.094	.091	.086	.084
		Std	.024	.016	.021	.015	.020	.014
	LR	Mean	.092	.087	.092	.085	.094	.090
		Std	.022	.014	.022	.015	.020	.016
	KM	Mean	.303	.321	.323	.328	.338	.352
		Std	.092	.083	.091	.083	.090	.081

Covariance matrices = 1:4 and Group separation = 2.2

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.102	.103	.104	.100	.100	.099
		Std	.022	.014	.021	.015	.021	.015
	LDF	Mean	.111	.111	.112	.107	.108	.106
		Std	.022	.016	.022	.015	.022	.015
	LR	Mean	.116	.114	.117	.110	.112	.109
		Std	.023	.016	.022	.015	.023	.016
	KM	Mean	.507	.498	.503	.500	.502	.498
		Std	.044	.030	.043	.031	.040	.033
0.25 : 0.75	LPM	Mean	.187	.188	.205	.204	.232	.233
		Std	.028	.020	.029	.022	.030	.024
	LDF	Mean	.185	.186	.197	.193	.211	.209
		Std	.029	.020	.027	.022	.030	.024
	LR	Mean	.187	.186	.197	.193	.205	.202
		Std	.028	.020	.027	.021	.029	.023
	KM	Mean	.486	.486	.488	.490	.496	.493
		Std	.044	.034	.044	.032	.041	.031
0.50 : 0.50	LPM	Mean	.213	.206	.209	.202	.232	.225
		Std	.032	.022	.029	.021	.034	.023
	LDF	Mean	.213	.206	.209	.202	.231	.224
		Std	.032	.022	.029	.021	.034	.023
	LR	Mean	.210	.203	.214	.208	.234	.235
		Std	.032	.020	.029	.021	.033	.022
	KM	Mean	.451	.468	.459	.467	.463	.470
		Std	.062	.048	.063	.051	.052	.039

Covariance matrices = 1:4 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.099	.101	.101	.101	.096	.101
		Std	.021	.015	.021	.015	.021	.015
	LDF	Mean	.110	.107	.107	.106	.100	.103
		Std	.023	.016	.022	.015	.022	.016
	LR	Mean	.117	.112	.112	.108	.104	.105
		Std	.024	.018	.024	.017	.023	.016
	KM	Mean	.503	.505	.500	.503	.507	.504
		Std	.041	.030	.041	.030	.043	.030
0.25 : 0.75	LPM	Mean	.290	.288	.283	.275	.264	.261
		Std	.032	.024	.030	.025	.030	.022
	LDF	Mean	.293	.290	.291	.283	.277	.274
		Std	.031	.024	.029	.024	.030	.022
	LR	Mean	.295	.290	.293	.285	.281	.278
		Std	.031	.023	.028	.024	.033	.021
	KM	Mean	.500	.502	.501	.498	.504	.501
		Std	.037	.027	.037	.028	.036	.026
0.50 : 0.50	LPM	Mean	.301	.299	.318	.311	.395	.394
		Std	.030	.023	.034	.024	.042	.030
	LDF	Mean	.301	.299	.318	.311	.394	.392
		Std	.030	.023	.034	.024	.042	.030
	LR	Mean	.301	.298	.323	.316	.393	.391
		Std	.029	.023	.034	.024	.041	.029
	KM	Mean	.484	.483	.481	.481	.485	.487
		Std	.040	.032	.044	.035	.047	.032

Covariance matrices = 4:1 and Group separation = 6.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.056	.055	.062	.060	.072	.070
		Std	.016	.012	.017	.013	.020	.014
	LDF	Mean	.052	.049	.053	.050	.057	.051
		Std	.015	.011	.016	.011	.018	.011
	LR	Mean	.056	.051	.058	.052	.061	.053
		Std	.016	.011	.017	.012	.019	.012
	KM	Mean	.444	.470	.456	.472	.463	.477
		Std	.088	.044	.067	.049	.069	.037
0.25 : 0.75	LPM	Mean	.083	.080	.085	.084	.092	.092
		Std	.020	.014	.019	.015	.022	.016
	LDF	Mean	.082	.079	.084	.083	.087	.086
		Std	.020	.014	.018	.015	.022	.015
	LR	Mean	.087	.081	.086	.082	.088	.084
		Std	.022	.014	.020	.015	.021	.014
	KM	Mean	.324	.400	.345	.389	.360	.423
		Std	.117	.087	.124	.104	.123	.096
0.50 : 0.50	LPM	Mean	.086	.083	.092	.091	.101	.100
		Std	.020	.015	.022	.016	.022	.016
	LDF	Mean	.086	.083	.092	.091	.101	.099
		Std	.019	.015	.022	.016	.022	.016
	LR	Mean	.095	.089	.091	.087	.090	.086
		Std	.021	.015	.022	.015	.020	.014
	KM	Mean	.334	.363	.325	.337	.297	.307
		Std	.087	.072	.086	.080	.088	.086

Covariance matrices = 4:1 and Group separation = 2.2

Population Proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.082	.085	.087	.086	.092	.093
		Std	.018	.015	.021	.013	.021	.014
	LDF	Mean	.080	.076	.079	.075	.079	.076
		Std	.018	.014	.019	.013	.019	.014
	LR	Mean	.081	.076	.082	.075	.081	.078
		Std	.018	.013	.020	.013	.019	.014
	KM	Mean	.470	.479	.473	.477	.472	.481
		Std	.056	.036	.052	.041	.053	.040
0.25 : 0.75	LPM	Mean	.154	.148	.153	.151	.157	.158
		Std	.025	.016	.027	.019	.026	.020
	LDF	Mean	.155	.149	.152	.149	.151	.150
		Std	.025	.016	.026	.018	.025	.019
	LR	Mean	.160	.151	.154	.150	.152	.151
		Std	.026	.016	.026	.018	.026	.019
	KM	Mean	.452	.467	.455	.470	.465	.475
		Std	.066	.051	.059	.045	.056	.042
0.50 : 0.50	LPM	Mean	.229	.227	.209	.204	.212	.209
		Std	.030	.023	.029	.018	.031	.021
	LDF	Mean	.228	.227	.209	.204	.211	.208
		Std	.030	.023	.029	.018	.031	.021
	LR	Mean	.240	.237	.214	.211	.208	.204
		Std	.031	.023	.030	.019	.030	.021
	KM	Mean	.462	.465	.458	.469	.463	.463
		Std	.059	.046	.062	.044	.053	.047

Covariance matrices = 4:1 and Group separation = 0.7

Population proportions	Method		Sample representativeness					
			Over		Equal		Under	
			Sample size		Sample size		Sample size	
			200	400	200	400	200	400
0.10 : 0.90	LPM	Mean	.094	.093	.095	.099	.098	.099
		Std	.023	.014	.019	.015	.022	.016
	LDF	Mean	.090	.086	.088	.090	.090	.090
		Std	.022	.014	.019	.015	.020	.015
	LR	Mean	.091	.087	.089	.091	.092	.091
		Std	.022	.014	.019	.015	.022	.015
	KM	Mean	.472	.482	.480	.482	.475	.489
		Std	.052	.036	.052	.039	.052	.037
0.25 : 0.75	LPM	Mean	.193	.190	.199	.194	.204	.207
		Std	.026	.020	.028	.021	.031	.024
	LDF	Mean	.194	.189	.198	.189	.196	.194
		Std	.026	.020	.029	.020	.029	.023
	LR	Mean	.196	.190	.198	.190	.196	.196
		Std	.027	.020	.029	.020	.030	.023
	KM	Mean	.466	.477	.473	.181	.465	.484
		Std	.054	.042	.054	.042	.050	.038
0.50 : 0.50	LPM	Mean	.389	.391	.319	.313	.309	.299
		Std	.041	.032	.035	.022	.033	.022
	LDF	Mean	.388	.389	.319	.313	.309	.299
		Std	.041	.032	.035	.022	.034	.022
	LR	Mean	.387	.388	.324	.318	.309	.299
		Std	.040	.031	.037	.022	.033	.023
	KM	Mean	.483	.486	.479	.487	.478	.484
		Std	.045	.036	.047	.031	.044	.033

Table T-3

ANOVA Results, Eta-squared, and Partial-Omega squared of Total Error Rate for Data Pattern I on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	105.67	52.84	24782.80	<.0001	.0668	.2766
Equality of covariance (COV)	2	19.44	9.72	4560.38	<.0001	.0123	.0657
Group separation (GS)	2	600.81	300.41	140908.0	<.0001	.3797	.6850
Sample representativeness (SR)	2	1.00	0.50	235.03	<.0001	.0006	.0036
Sample size (SS)	1	0.014	0.014	6.66	.0099	<.0001	<.0001
PP*COV	4	10.11	2.53	1185.34	<.0001	.0064	.0353
PP*GS	4	84.08	21.02	9859.56	<.0001	.0531	.2333
PP*SR	4	0.96	0.24	112.89	<.0001	.0006	.0034
PP*SS	2	0.016	0.008	3.82	.0220	<.0001	<.0001
COV*GS	4	2.18	0.54	255.43	<.0001	.0014	.0078
COV*SR	4	0.95	0.24	111.58	<.0001	.0006	.0034
COV*SS	2	0.0009	0.0004	0.20	.8165	<.0001	<.0001
GS*SR	4	0.36	0.09	42.38	<.0001	.0002	.0013
GS*SS	2	0.0052	0.0026	1.22	.2960	<.0001	<.0001
SR*SS	2	0.008	0.004	1.87	.1538	<.0001	<.0001
Error (between)	32358	68.99	0.0021				
Within-subject effects:							
Method (4M)	3	292.53	97.51	208717.0	<.0001	.1849	.8285
4M*PP	6	289.03	48.17	103109.0	<.0001	.1827	.8268
4M*COV	6	7.22	1.20	2576.55	<.0001	.0046	.1065
4M*GS	6	14.66	2.44	5230.26	<.0001	.0093	.1949
4M*SR	6	0.68	0.11	241.15	<.0001	.0004	.0110
4M*SS	3	0.043	0.014	30.47	<.0001	<.0001	.0007
4M*PP*COV	12	3.71	0.31	661.45	<.0001	.0023	.0576
4M*PP*GS	12	29.25	2.44	5217.19	<.0001	.0185	.3257
4M*PP*SR	12	0.64	0.05	114.92	<.0001	.0004	.0104
4M*PP*SS	6	0.022	0.0036	7.66	<.0001	<.0001	.0003
4M*COV*GS	12	3.78	0.32	675.07	<.0001	.0024	.0587
4M*COV*SR	12	0.63	0.052	111.73	<.0001	.0004	.0101
4M*COV*SS	6	0.0007	0.0001	0.27	.9522	<.0001	<.0001
4M*GS*SR	12	0.22	0.018	39.12	<.0001	.0001	.0035
4M*GS*SS	6	0.0007	0.0001	0.24	.9622	<.0001	<.0001
4M*SR*SS	6	0.0064	0.0011	2.28	.0337	<.0001	.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	45.35	0.0005				

Note. Eta square (η^2) is defined as: $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table T-4

ANOVA Results, Eta-squared, and Partial-Omega squared of Total Error Rate for Data Pattern II on Comparing Four Methods

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Between-subject effects:							
Population proportion (PP)	2	180.04	90.02	40628.50	<.0001	.0537	.3854
Equality of covariance (COV)	2	12.02	6.01	2713.00	<.0001	.0036	.0402
Group separation (GS)	2	375.38	187.69	84709.80	<.0001	.1119	.5666
Sample representativeness (SR)	2	0.57	0.29	128.63	<.0001	.0002	.0020
Sample size (SS)	1	0.011	0.011	4.91	.0268	<.0001	<.0001
PP*COV	4	7.48	1.87	843.86	<.0001	.0022	.0254
PP*GS	4	119.79	29.95	13516.10	<.0001	.0357	.2943
PP*SR	4	0.52	0.13	58.85	<.0001	.0002	.0018
PP*SS	2	0.02	0.01	4.42	.0121	<.0001	.0001
COV*GS	4	3.30	0.83	372.63	<.0001	.0010	.0113
COV*SR	4	1.16	0.29	131.13	<.0001	.0003	.0040
COV*SS	2	0.10	0.05	22.59	<.0001	<.0001	.0003
GS*SR	4	0.23	0.06	25.87	<.0001	.0001	.0008
GS*SS	2	0.077	0.039	17.40	<.0001	<.0001	.0003
SR*SS	2	0.011	0.006	2.49	.0831	<.0001	<.0001
Error (between)	32358	71.69	0.0022				
Within-subject effects:							
Method (4M)	3	2297.18	765.73	778723.0	<.0001	.6848	.9474
4M*PP	6	145.59	24.27	24677.50	<.0001	.0434	.5332
4M*COV	6	1.45	0.24	244.97	<.0001	.0004	.0112
4M*GS	6	26.65	4.44	4516.30	<.0001	.0079	.1729
4M*SR	6	0.22	0.037	37.17	<.0001	.0001	.0017
4M*SS	3	0.86	0.29	292.84	<.0001	.0003	.0067
4M*PP*COV	12	0.41	0.034	34.57	<.0001	.0001	.0031
4M*PP*GS	12	8.84	0.74	749.28	<.0001	.0026	.0648
4M*PP*SR	12	0.31	0.026	26.19	<.0001	.0001	.0023
4M*PP*SS	6	0.12	0.02	20.35	<.0001	<.0001	.0009
4M*COV*GS	12	4.12	0.34	349.54	<.0001	.0012	.0313
4M*COV*SR	12	0.19	0.016	16.07	<.0001	.0001	.0014
4M*COV*SS	6	0.20	0.033	33.98	<.0001	.0001	.0015
4M*GS*SR	12	0.11	0.0089	9.02	<.0001	<.0001	.0007
4M*GS*SS	6	0.28	0.047	48.11	<.0001	.0001	.0022
4M*SR*SS	6	0.016	0.0026	2.63	.0151	<.0001	.0001

Source of Variation	df	SS	MS	F	p	Eta-Squared	Partial Omega-Squared
Error (within)	97074	95.45	0.0010				

Note. Eta square (η^2) is defined as: $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, where SS_{effect} is the effect sum of squares, and SS_{total} is the total sum of squares. Partial omega squared was calculated from the formula: $\omega^2_{partial} = \frac{df_{effect}(F_{effect} - 1)}{df_{effect}(F_{effect} - 1) + N}$, where df_{effect} is the degrees of freedom for the effect, F_{effect} is the F ratio for the effect, and N equals 200×3×3×3×3×2×4 (=129600) in this ANOVA model.

Table T-5

Means and Standard Deviations of Total Error Rate for Method by Population Proportion Interaction

	Method							
Population proportions	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
0.10:0.90	.088	.024	.078	.030	.079	.031	.338	.105
0.25:0.75	.165	.074	.162	.075	.162	.075	.241	.107
0.50:0.50	.217	.106	.217	.106	.217	.107	.212	.086
				Data Pattern II				
0.10:0.90	.088	.023	.080	.030	.082	.030	.489	.046
0.25:0.75	.167	.074	.165	.075	.166	.075	.466	.070
0.50:0.50	.221	.106	.221	.106	.221	.106	.438	.090

Table T-6

Means and Standard Deviations of Total Error Rate for Method by Equality of Covariance Matrices Interaction

Equality of covariance matrices	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I				
1:4	.162	.096	.158	.100	.159	.100	.285	.108
Equal	.162	.093	.158	.096	.159	.095	.277	.112
4:1	.146	.088	.140	.090	.141	.091	.228	.112
				Data Pattern II				
1:4	.164	.097	.161	.100	.162	.100	.467	.073
Equal	.164	.095	.161	.097	.163	.096	.481	.058
4:1	.148	.088	.143	.090	.145	.090	.445	.083

Table T-7

Means and Standard Deviations of Total Error Rate for Method by Group Separation Interaction

Group separation	Method							
	LPM		LDF		LR		KM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	Data Pattern I							
6.7	.080	.022	.071	.026	.070	.025	.151	.074
2.2	.165	.059	.161	.060	.162	.060	.279	.080
0.7	.225	.105	.224	.106	.225	.106	.361	.066
	Data Pattern II							
6.7	.082	.023	.074	.027	.075	.026	.420	.098
2.2	.167	.061	.164	.061	.166	.061	.482	.047
0.7	.226	.107	.227	.107	.228	.106	.491	.040

Table T-8

Means and Standard Deviations of Total Error Rate for Method by Population Proportions by Group Separation Interaction

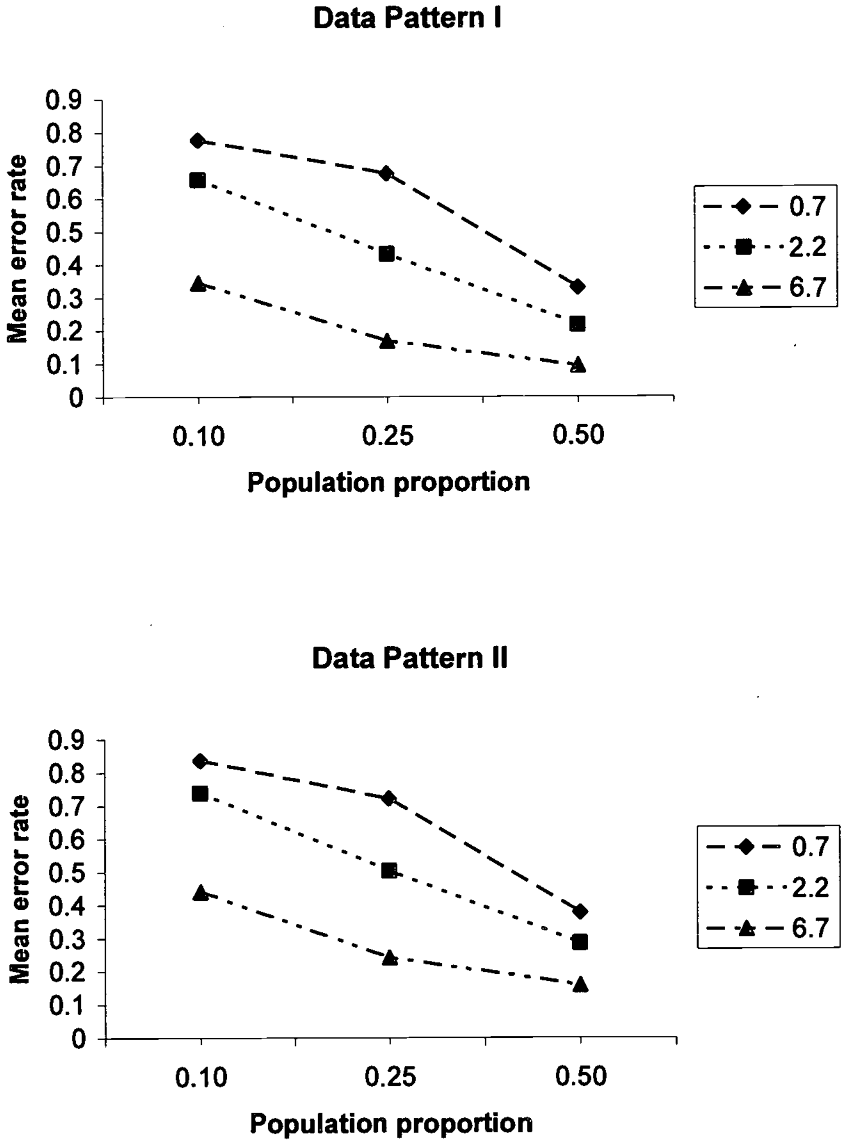
		Method							
Group separation	Population proportions	LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				Data Pattern I					
6.7	0.10:0.90	.069	.021	.045	.014	.046	.014	.224	.080
	0.25:0.75	.077	.019	.075	.019	.075	.019	.122	.043
	0.50:0.50	.093	.019	.093	.019	.091	.019	.108	.021
2.2	0.10:0.90	.096	.019	.091	.023	.092	.023	.364	.056
	0.25:0.75	.180	.036	.173	.032	.173	.031	.253	.062
	0.50:0.50	.219	.030	.219	.030	.222	.030	.219	.027
0.7	0.10:0.90	.099	.019	.098	.020	.099	.020	.426	.046
	0.25:0.75	.237	.041	.237	.045	.238	.045	.349	.050
	0.50:0.50	.339	.048	.338	.048	.339	.047	.308	.034
				Data Pattern II					
6.7	0.10:0.90	.069	.021	.047	.014	.051	.015	.482	.051
	0.25:0.75	.079	.020	.078	.019	.079	.020	.424	.092
	0.50:0.50	.097	.021	.097	.021	.095	.021	.354	.098
2.2	0.10:0.90	.095	.019	.093	.022	.095	.023	.491	.043
	0.25:0.75	.182	.035	.177	.031	.177	.030	.482	.046
	0.50:0.50	.224	.031	.223	.031	.227	.031	.472	.050
0.7	0.10:0.90	.099	.018	.099	.020	.100	.021	.494	.042
	0.25:0.75	.239	.042	.240	.046	.241	.046	.491	.040
	0.50:0.50	.342	.046	.342	.045	.342	.045	.487	.038

Table T-9

Means and Standard Deviations of Total Error Rate for Method by Population Proportion by Equality of Covariance Matrices Interaction

Population Proportion	Equality of covariance matrices	Method							
		LPM		LDF		LR		KM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Data Pattern I									
0.10:0.90	1:4	.093	.023	.083	.036	.085	.037	.368	.082
	Equal	.089	.024	.080	.028	.081	.028	.354	.086
	4:1	.083	.024	.070	.023	.072	.024	.292	.125
0.25:0.75	1:4	.181	.091	.179	.093	.179	.093	.279	.094
	Equal	.169	.068	.167	.068	.167	.068	.257	.105
	4:1	.144	.053	.139	.051	.140	.052	.187	.098
0.50:0.50	1:4	.212	.106	.211	.106	.211	.107	.207	.079
	Equal	.228	.105	.228	.105	.228	.105	.221	.100
	4:1	.212	.107	.212	.106	.212	.108	.206	.078
Data Pattern II									
0.10:0.90	1:4	.093	.023	.085	.036	.089	.037	.499	.038
	Equal	.088	.023	.082	.028	.083	.028	.495	.040
	4:1	.082	.023	.073	.022	.075	.022	.473	.053
0.25:0.75	1:4	.184	.092	.182	.093	.183	.093	.476	.054
	Equal	.171	.069	.170	.069	.171	.069	.483	.050
	4:1	.146	.051	.143	.050	.144	.050	.437	.090
0.50:0.50	1:4	.215	.106	.214	.105	.215	.107	.425	.094
	Equal	.233	.105	.233	.105	.234	.104	.464	.075
	4:1	.215	.106	.214	.105	.215	.107	.424	.093

Figure G1-1. Population proportion by group separation interaction on Group 1 error rate.



Data Pattern I

Data Pattern II

Figure G1-2. Equality of covariance matrices by group separation interaction on Group 1 error rate.

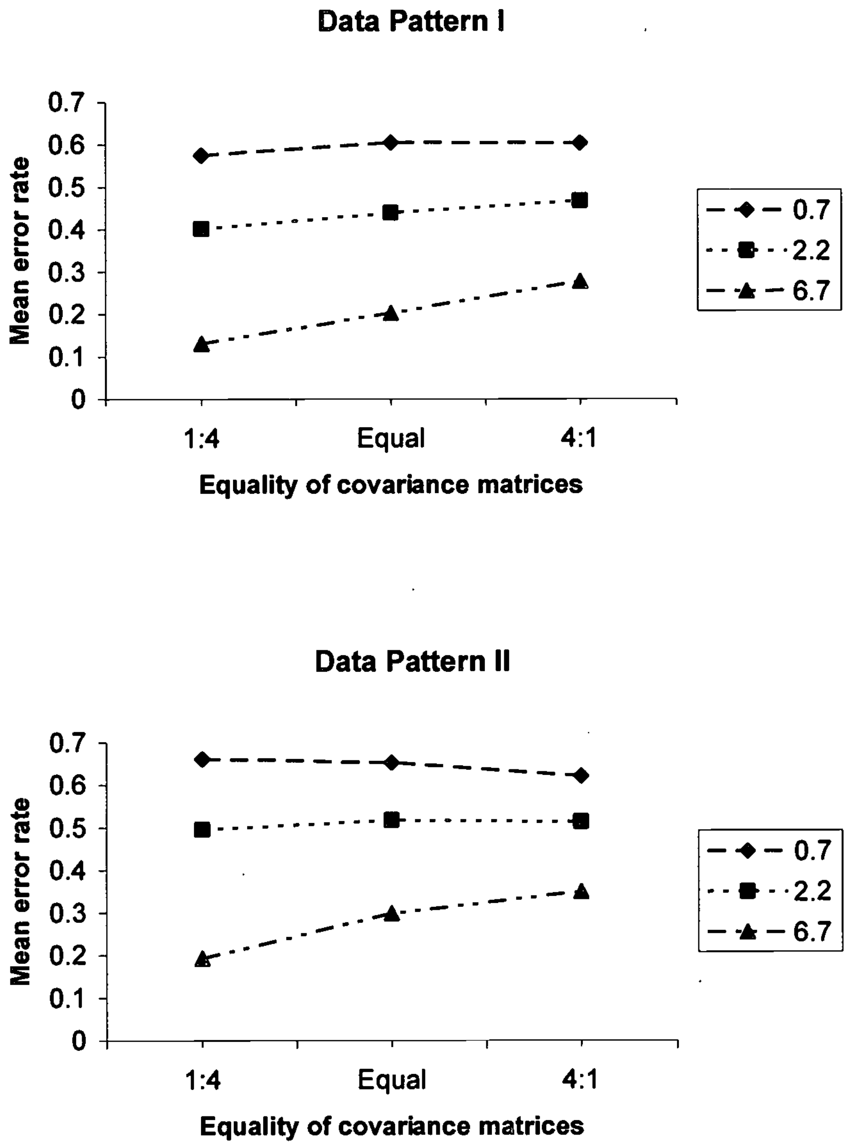
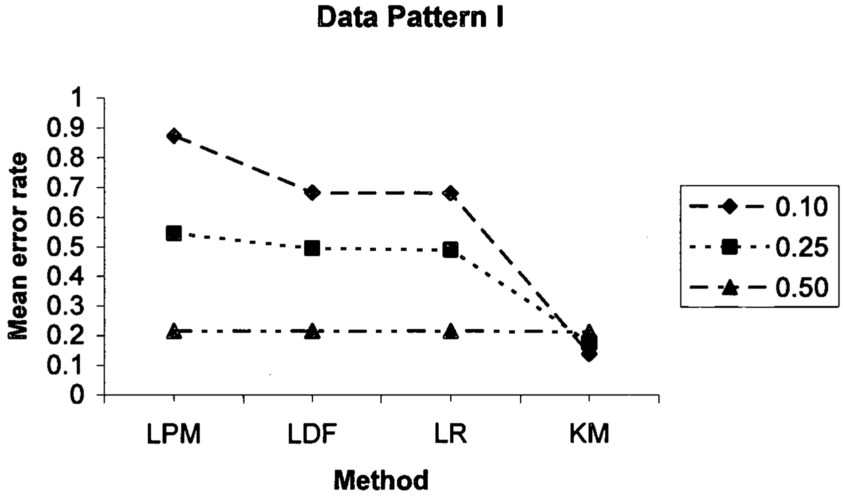


Figure G1-3. Method by population proportion interaction on Group 1 error rate.



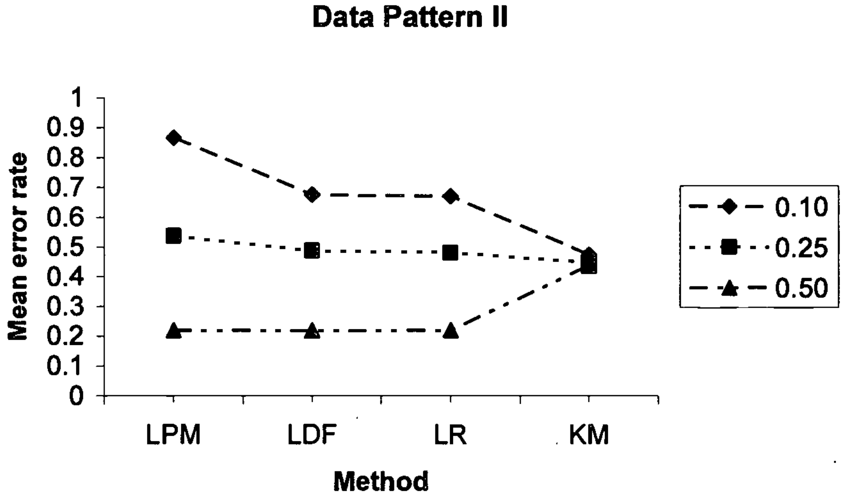
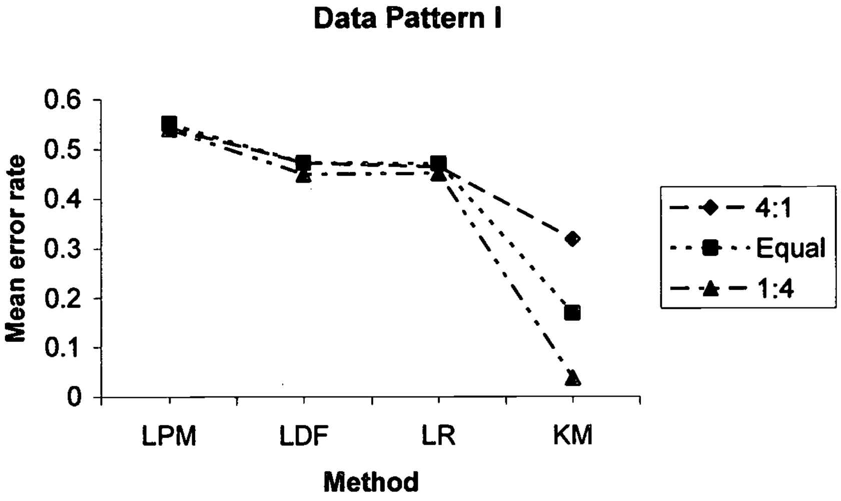


Figure G1-4. Method by equality of covariance matrices interaction on Group 1 error rate.



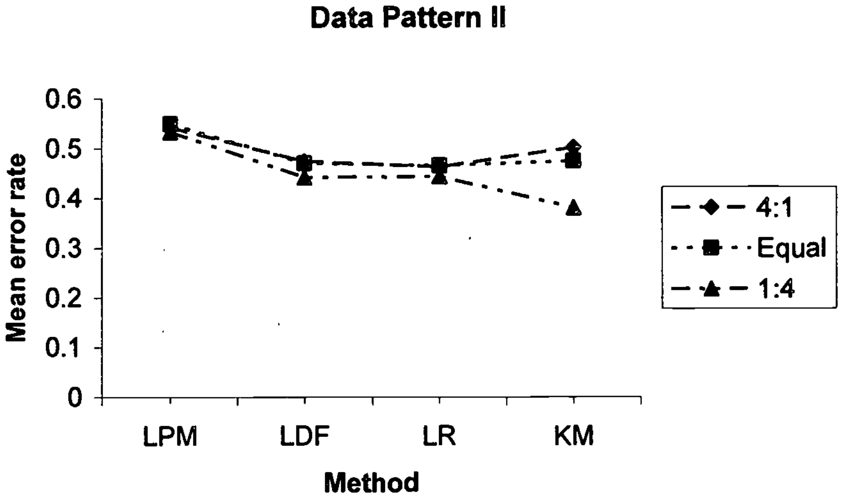


Figure G1-5. Method by group separation interaction on Group 1 error rate.

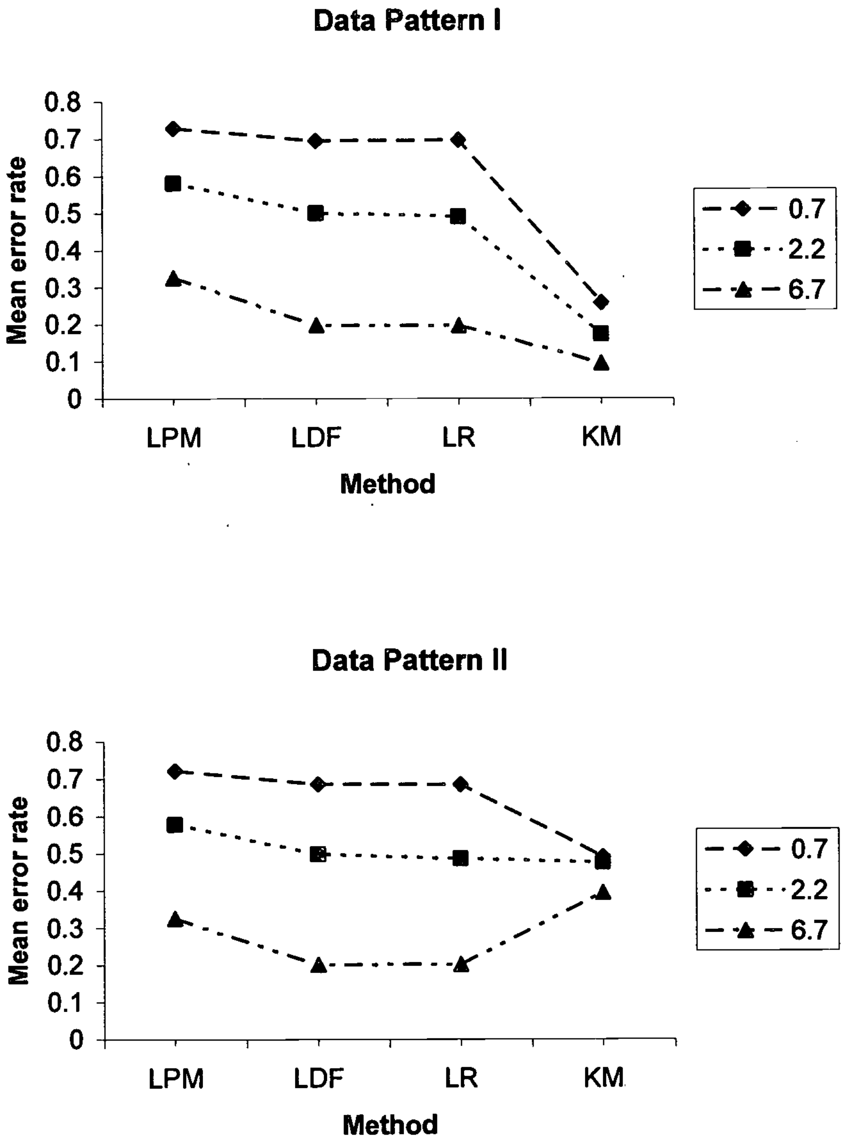
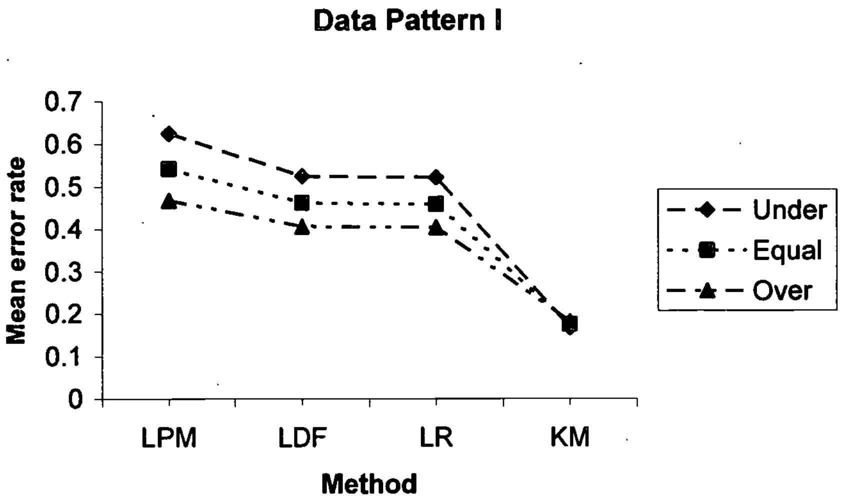


Figure G1-6. Method by sample representativeness interaction on Group 1 error rate.



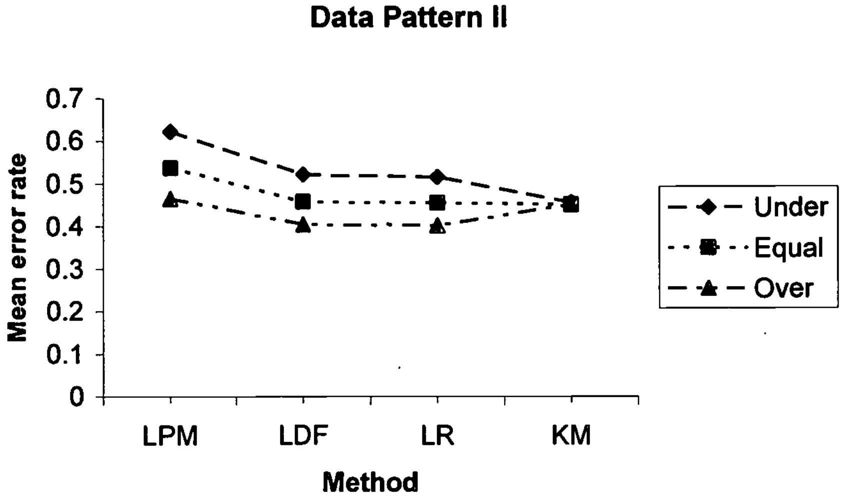
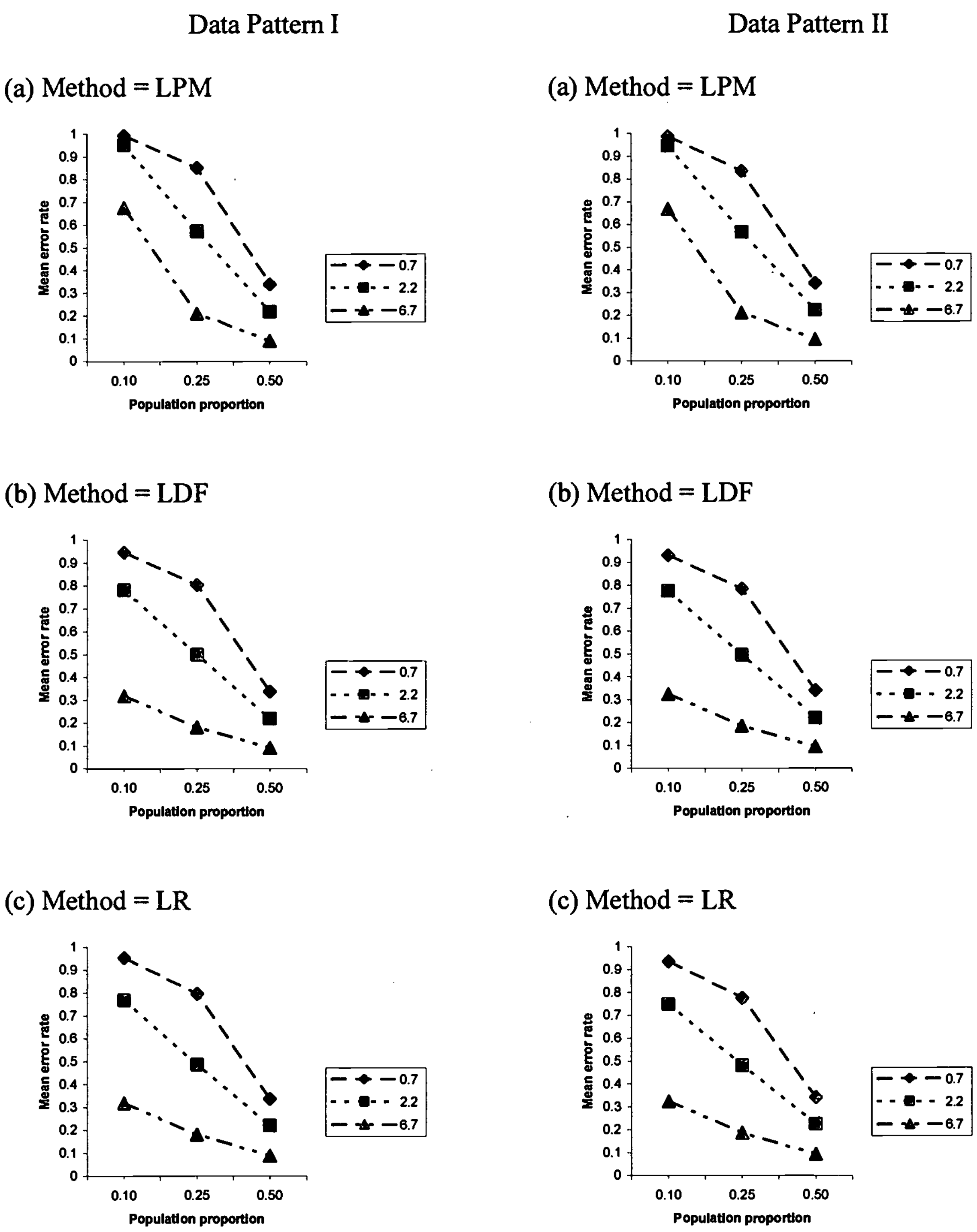


Figure G1-7. Method by population proportion by group separation interaction on Group 1 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

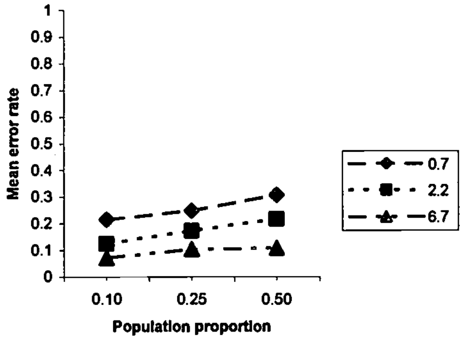
(b) Method = LDF

(b) Method = LDF

(c) Method = LR

(c) Method = LR

(d) Method = KM



(d) Method = KM

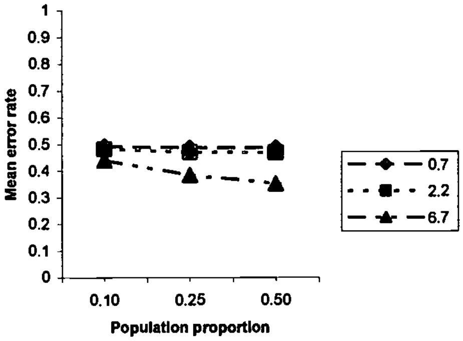
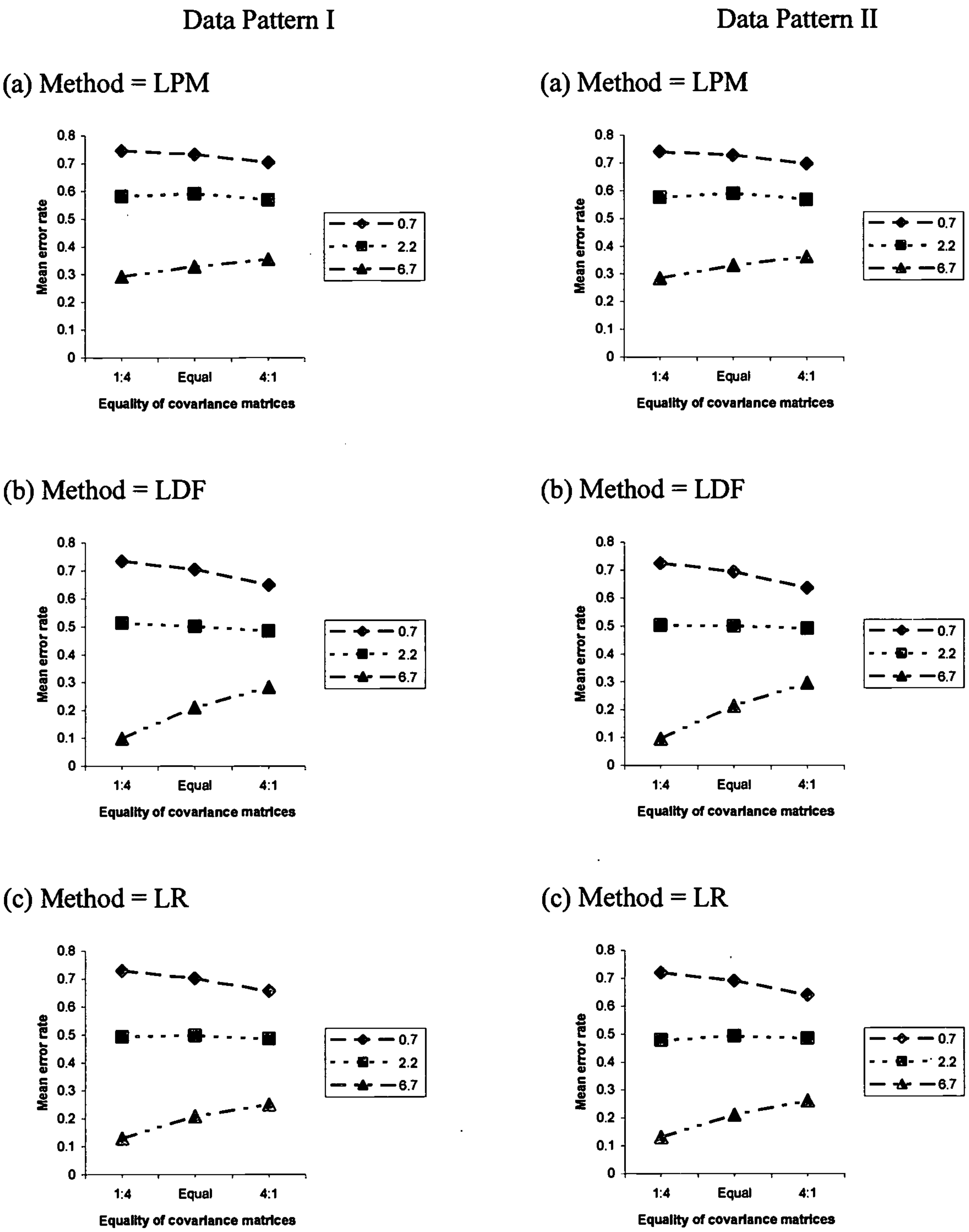


Figure G1-8. Method by equality of covariance matrices by group separation interaction on Group 1 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

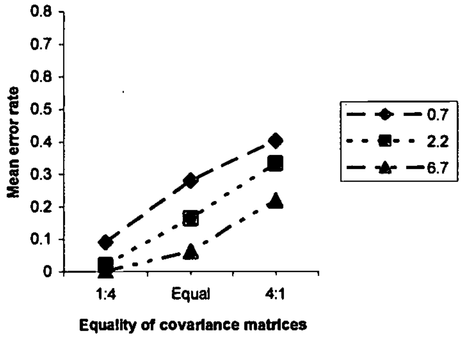
(b) Method = LDF

(b) Method = LDF

(c) Method = LR

(c) Method = LR

(d) Method = KM



(d) Method = KM

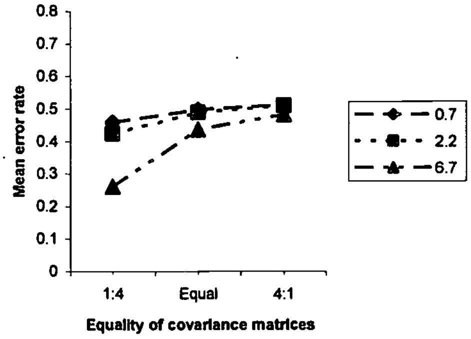


Figure G2-1. Population proportion by equality of covariance matrices interaction on Group 2 error rate.

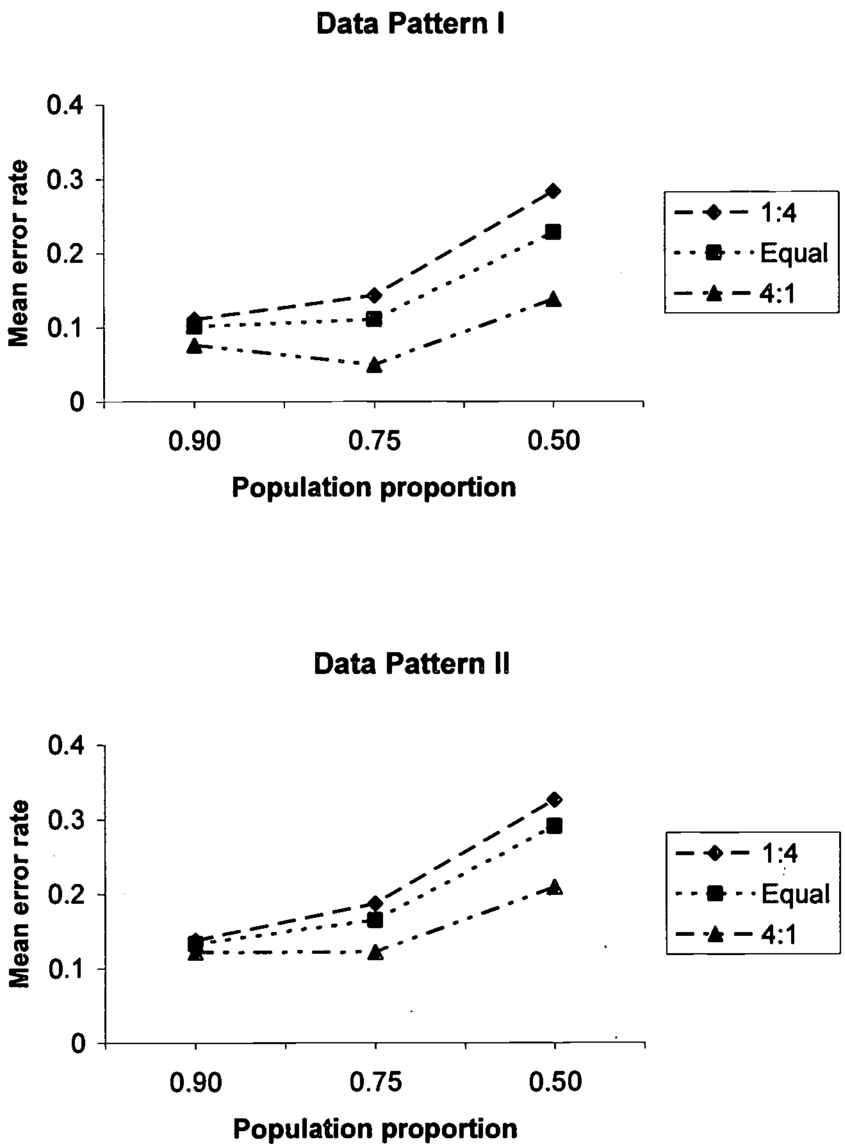


Figure G2-2. Population proportion by group separation interaction on Group 2 error rate.

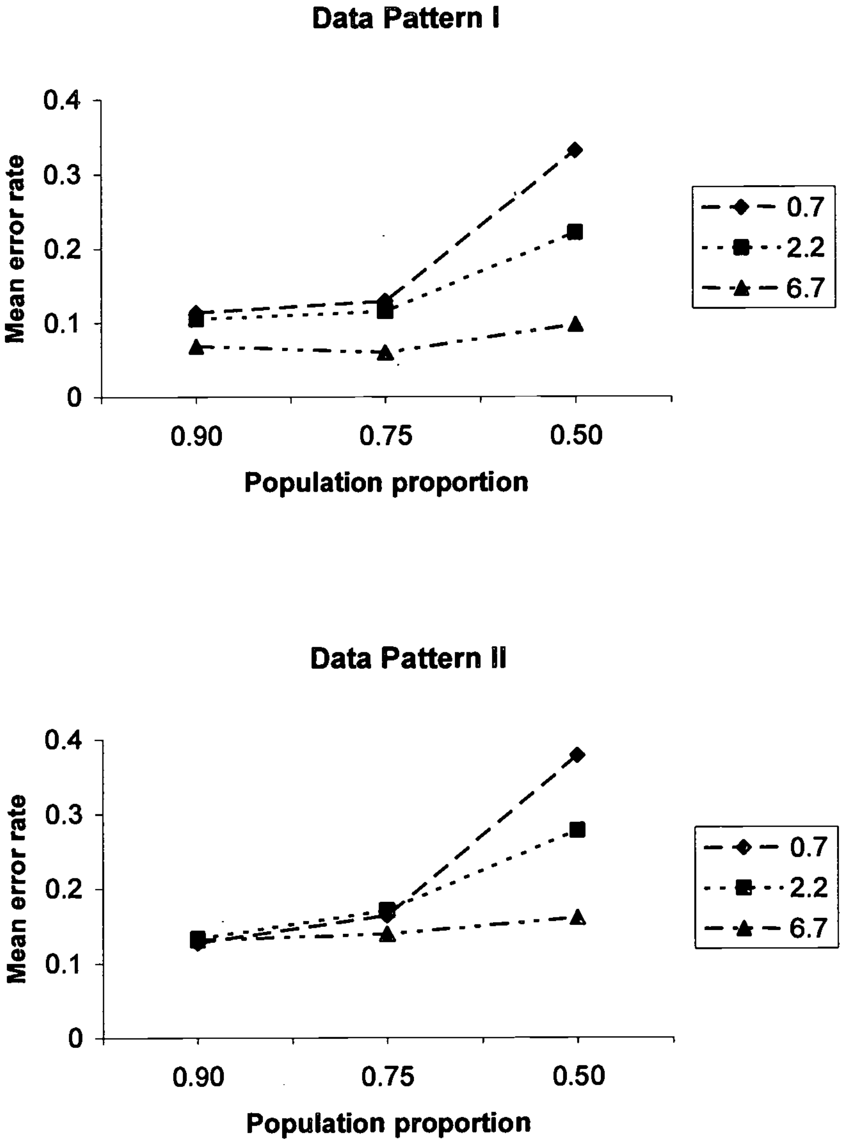
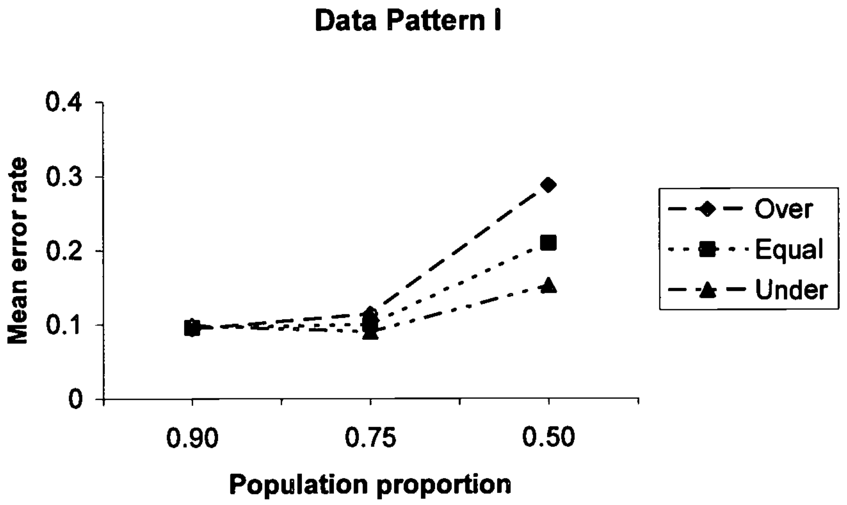


Figure G2-3. Population proportion by sample representativeness interaction on Group 2 error rate.



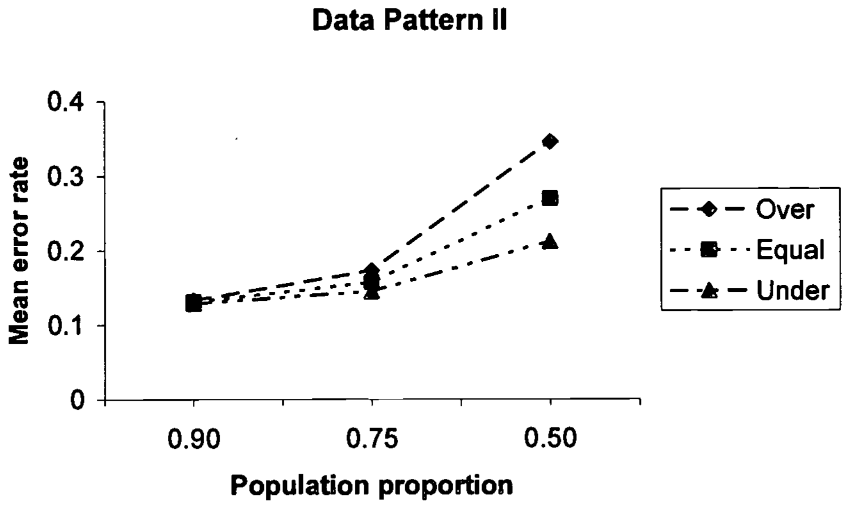
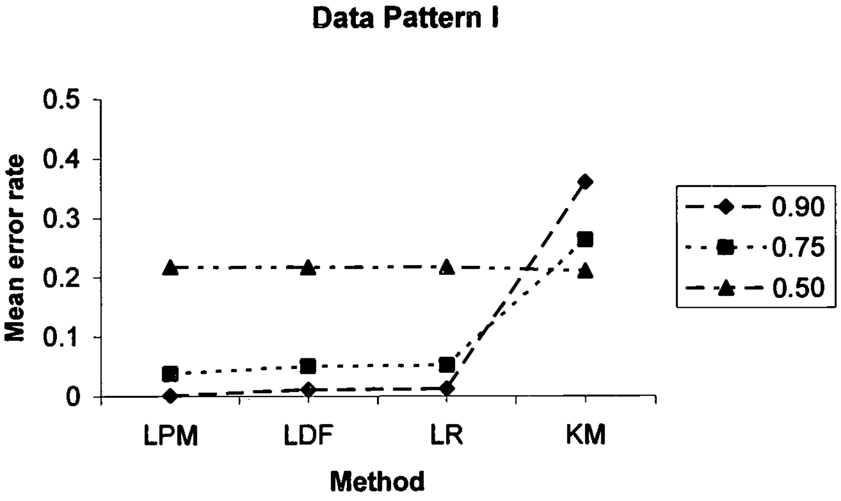


Figure G2-4. Method by population proportion interaction on Group 2 error rate.



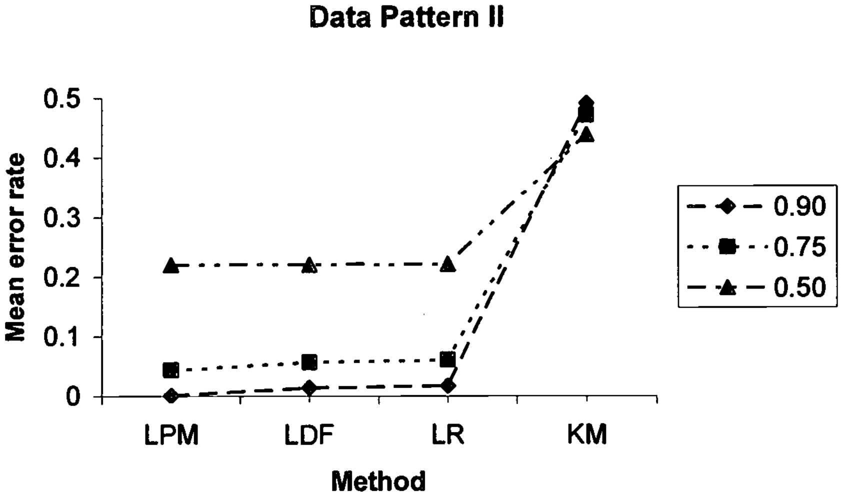
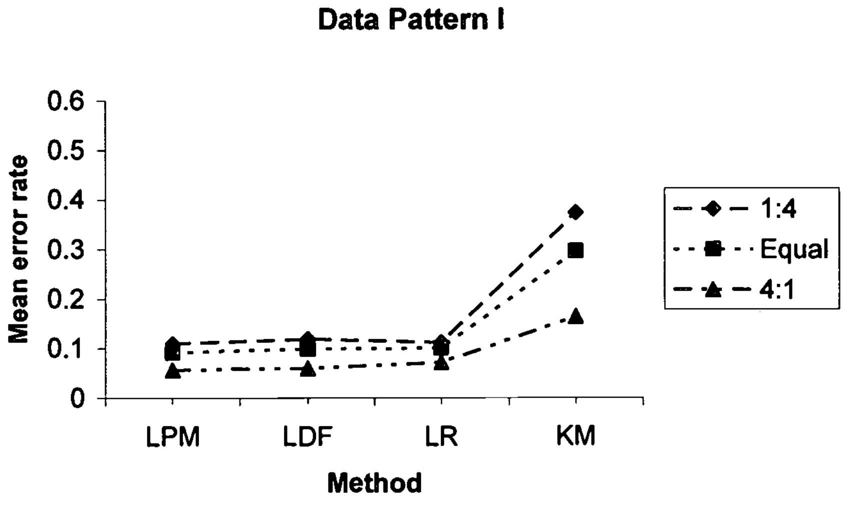


Figure G2-5. Method by equality of covariance matrices interaction on Group 2 error rate.



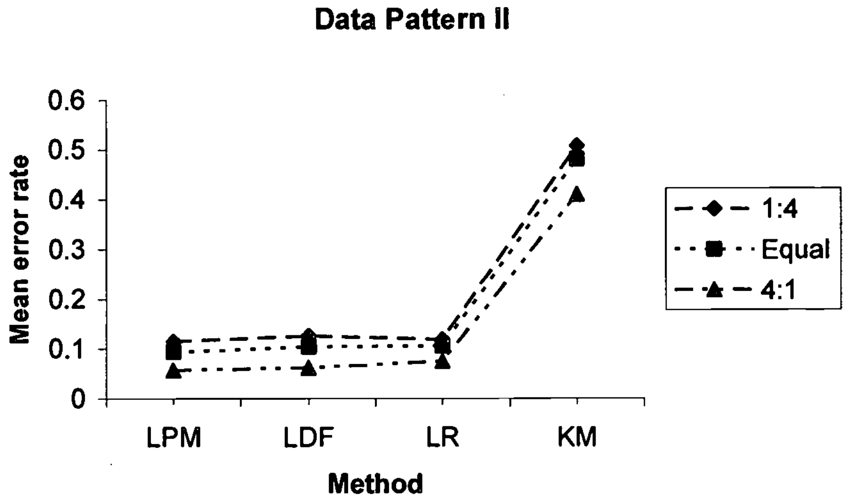
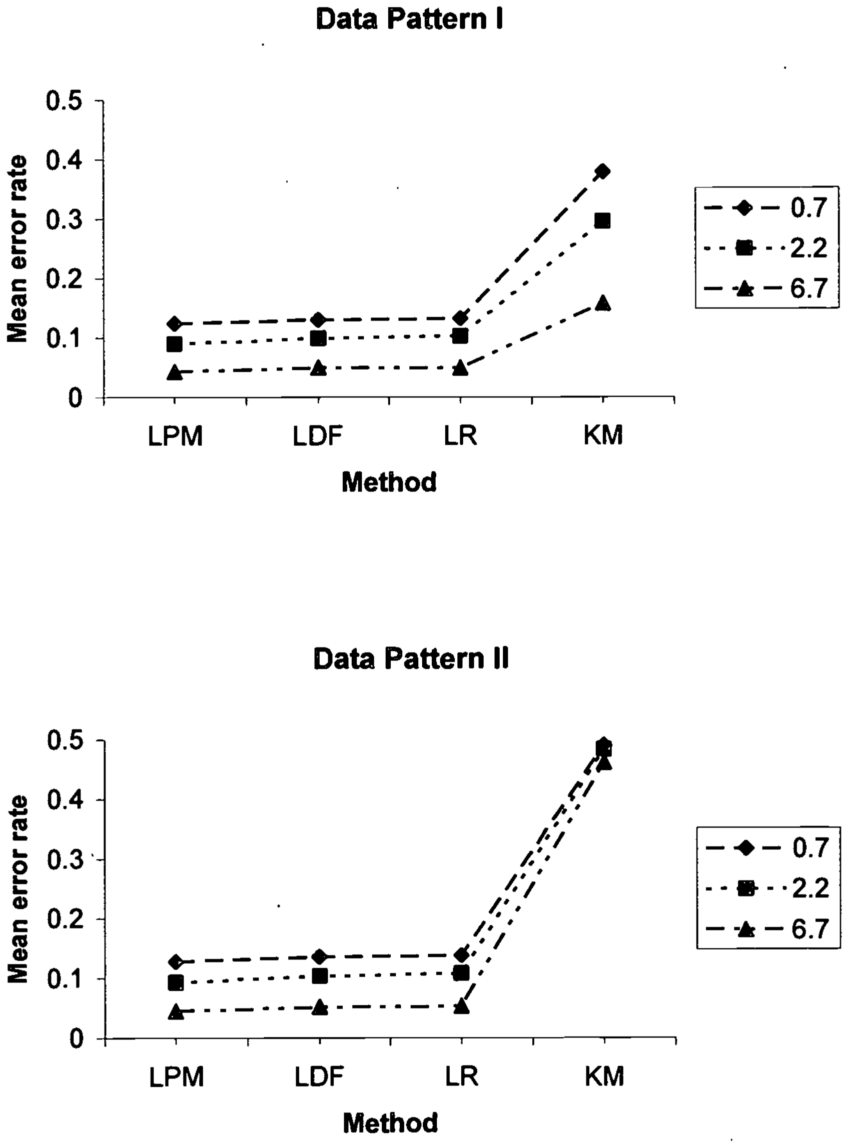


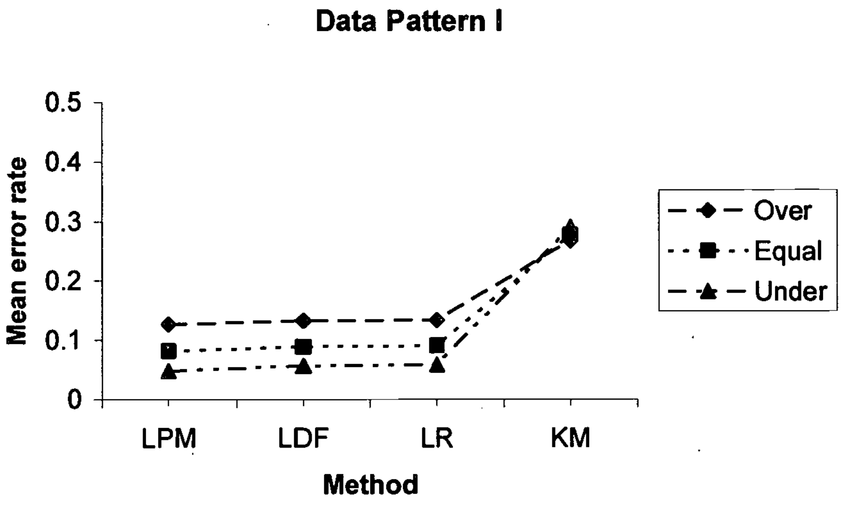
Figure G2-6. Method by group separation interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

Figure G2-7. Method by sample representativeness interaction on Group 2 error rate.



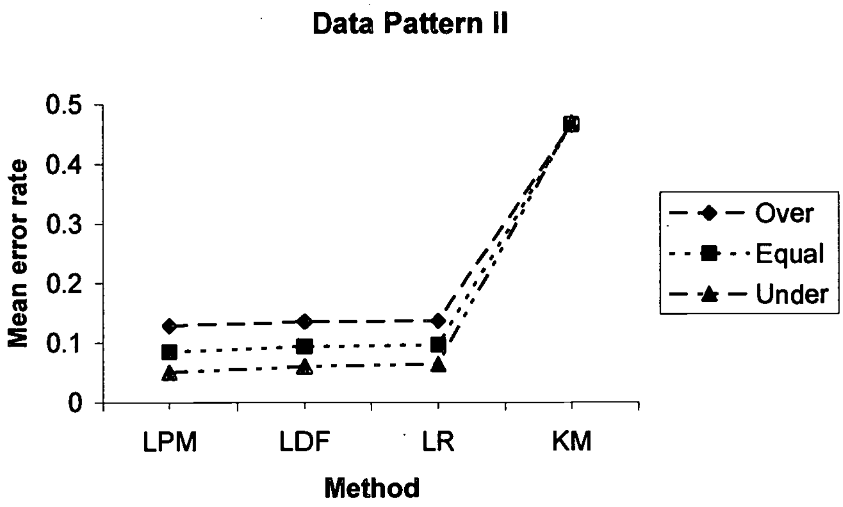
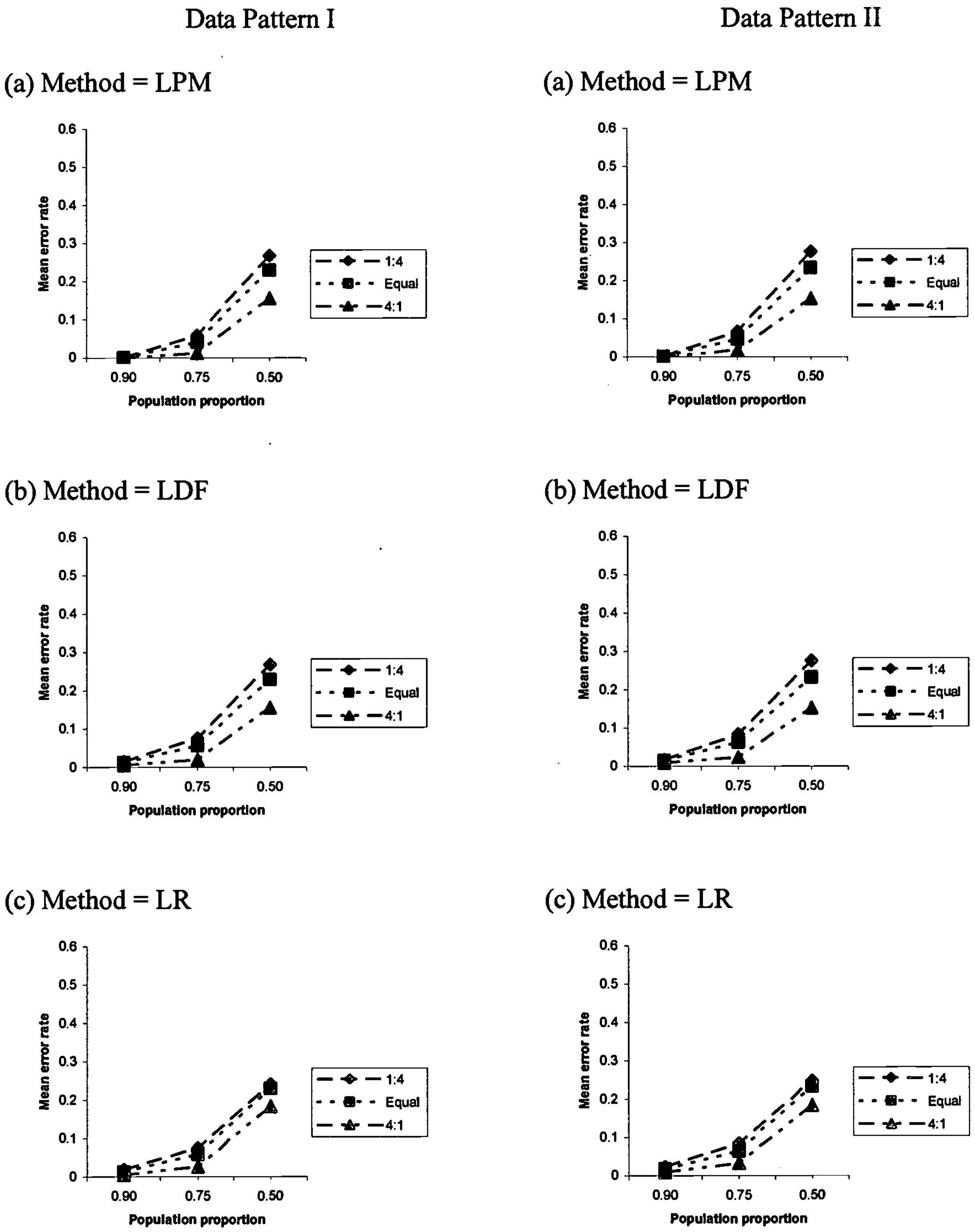


Figure G2-8. Method by population proportion by equality of covariance matrices interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

(b) Method = LDF

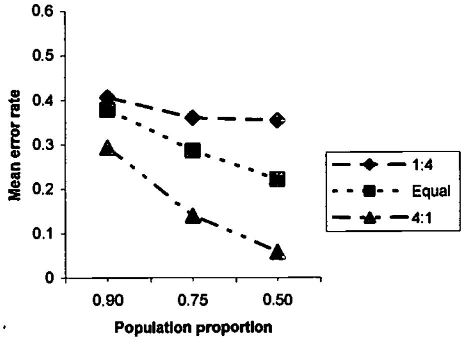
(b) Method = LDF

(c) Method = LR

(c) Method = LR

Data Pattern I

(d) Method = KM



Data Pattern II

(d) Method = KM

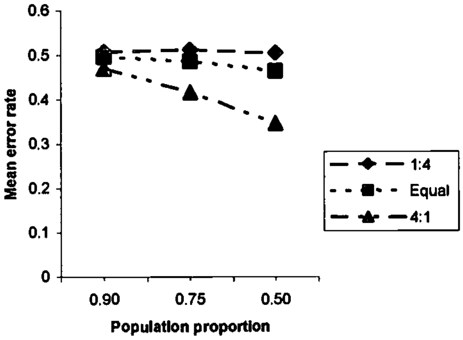
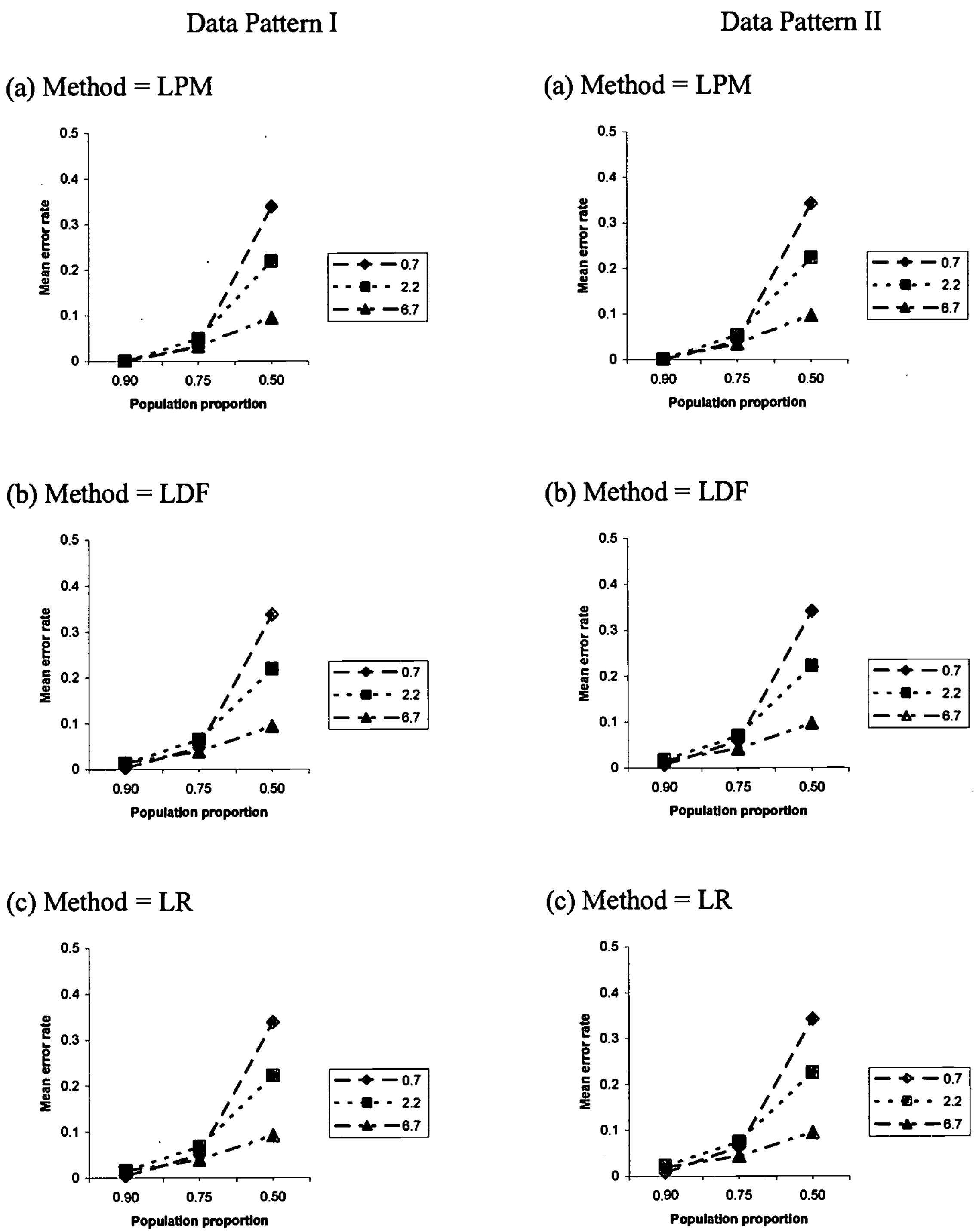


Figure G2-9. Method by population proportion by group separation interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

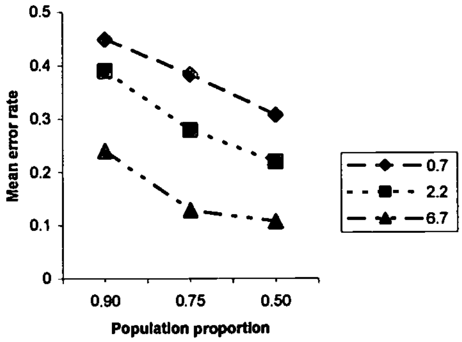
(a) Method = LPM

(b) Method = LDF

(c) Method = LR

Data Pattern I

(d) Method = KM



Data Pattern II

(d) Method = KM

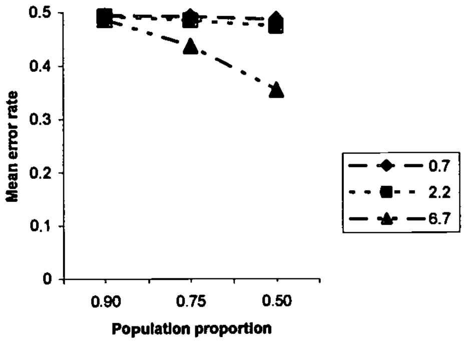
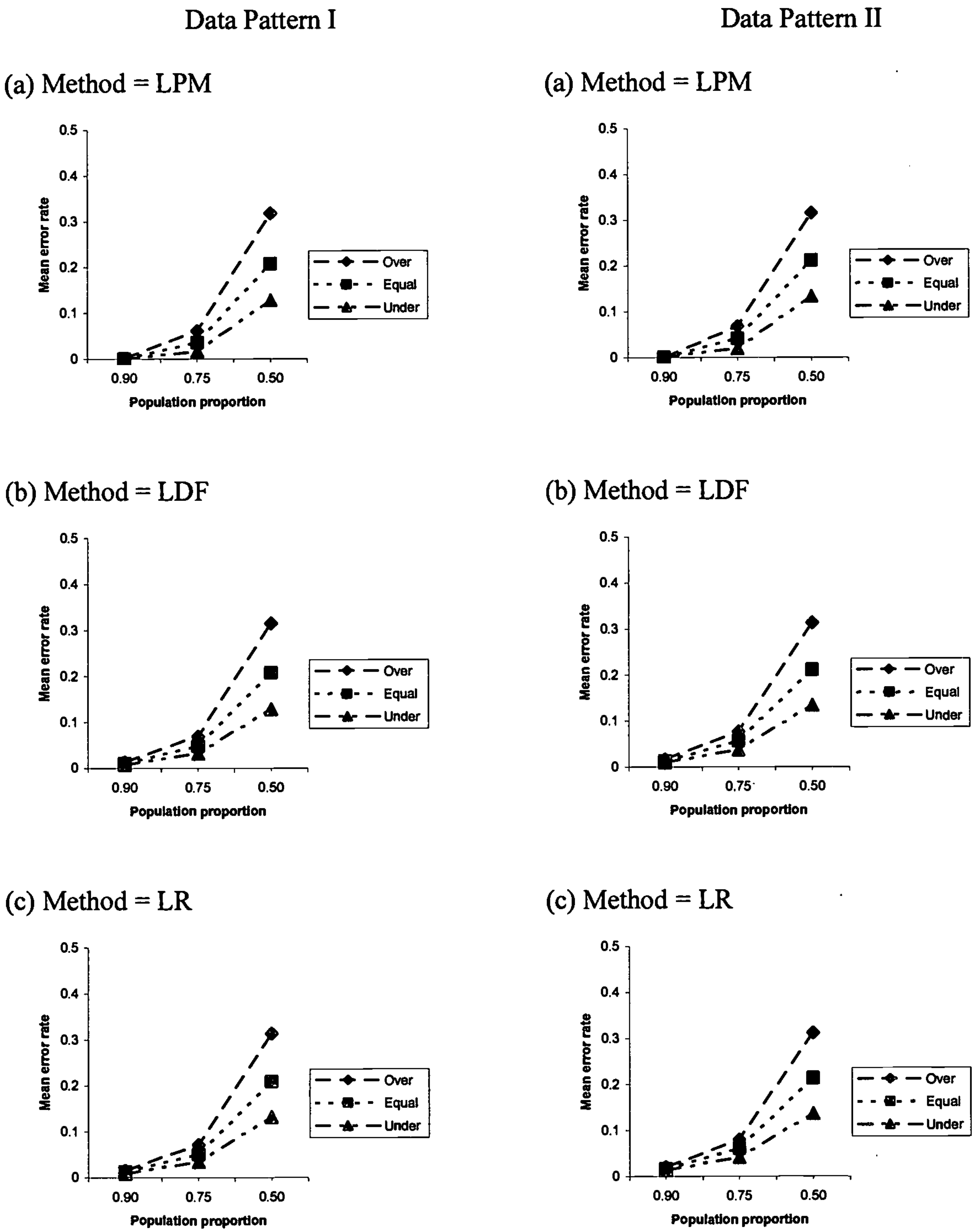


Figure G2-10. Method by population proportion by sample representativeness interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

(b) Method = LDF

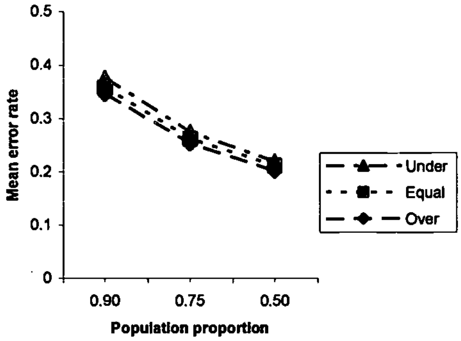
(b) Method = LDF

(c) Method = LR

(c) Method = LR

Data Pattern I

(d) Method = KM



Data Pattern II

(d) Method = KM

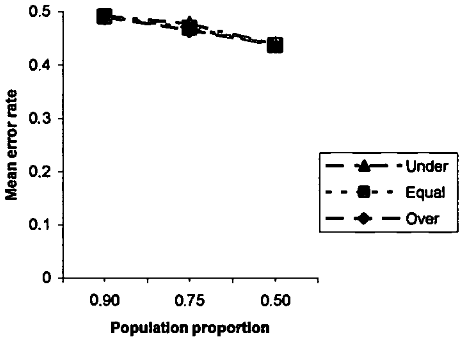
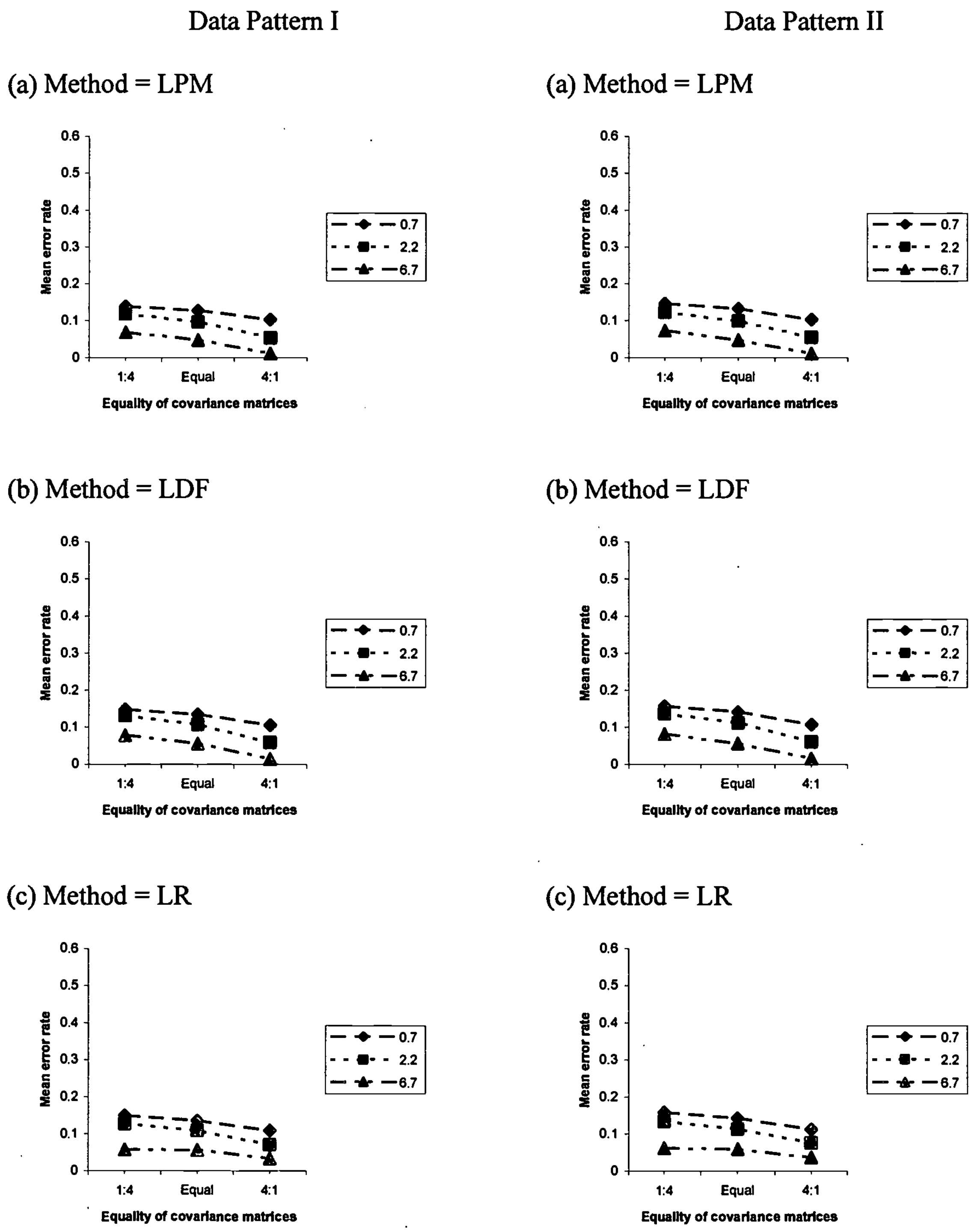


Figure G2-11. Method by equality of covariance matrices by group separation interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

(b) Method = LDF

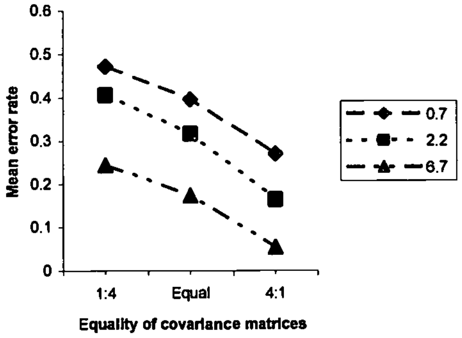
(b) Method = LDF

(c) Method = LR

(c) Method = LR

Data Pattern I

(d) Method = KM



Data Pattern II

(d) Method = KM

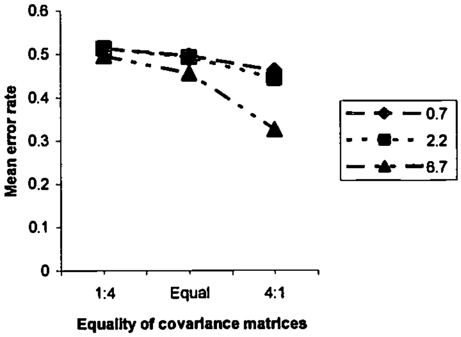
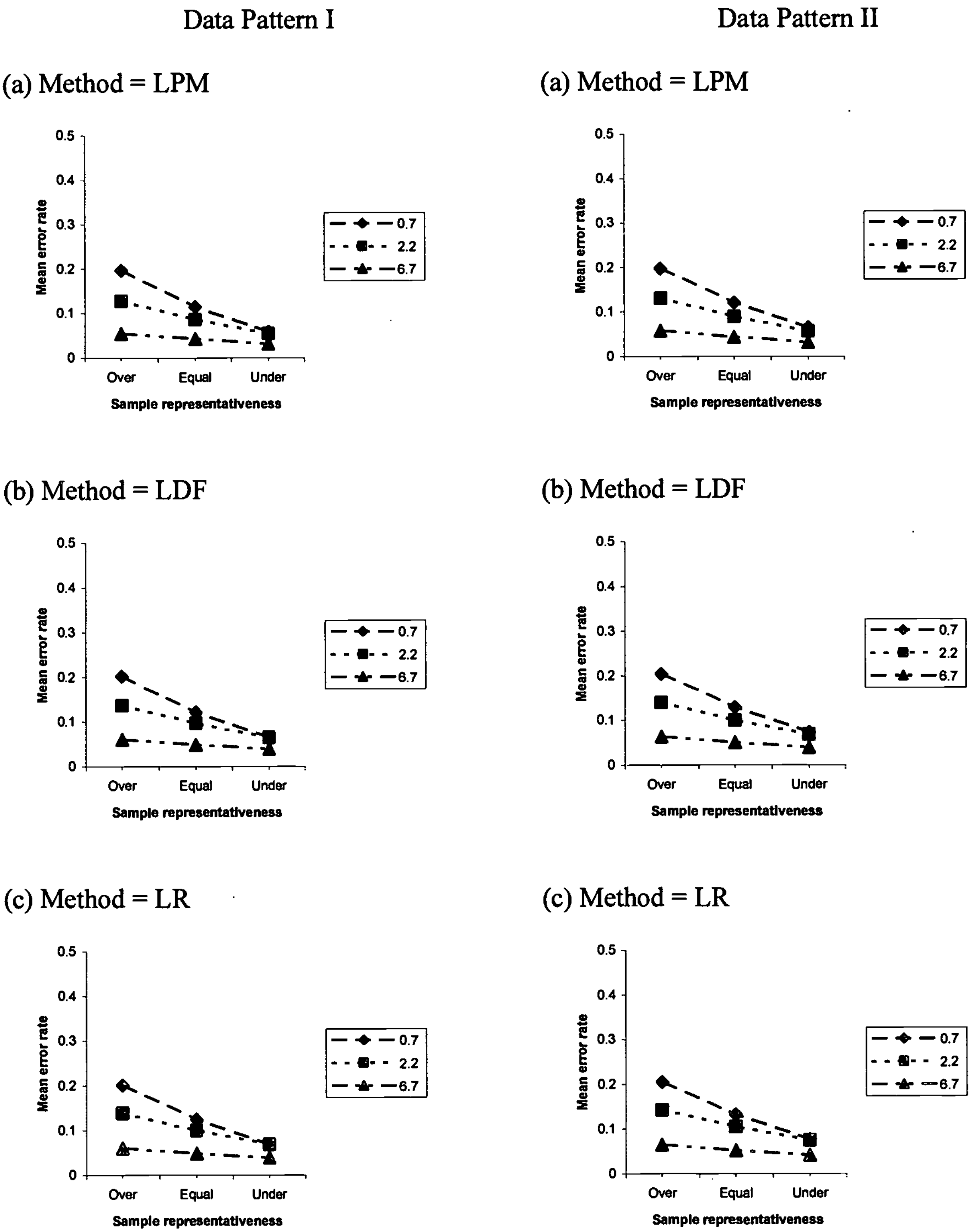


Figure G2-12. Method by group separation by sample representativeness interaction on Group 2 error rate.



Data Pattern I

Data Pattern II

(a) Method = LPM

(a) Method = LPM

(b) Method = LDF

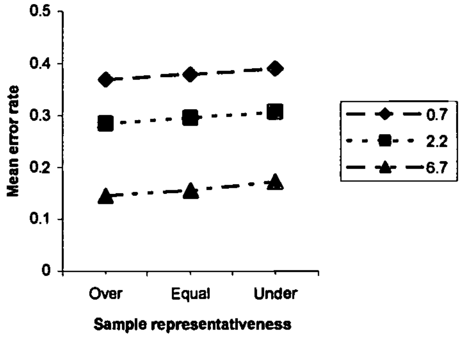
(b) Method = LDF

(c) Method = LR

(c) Method = LR

Data Pattern I

(d) Method = KM



Data Pattern II

(d) Method = KM

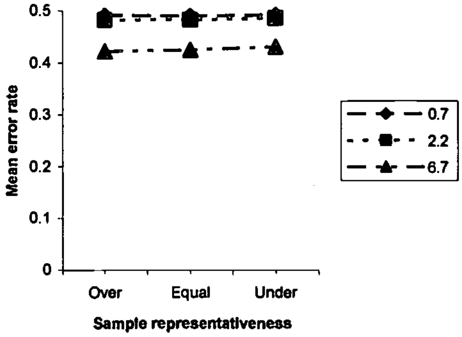


Figure T-1. Population proportion by group separation interaction on Total error rate.

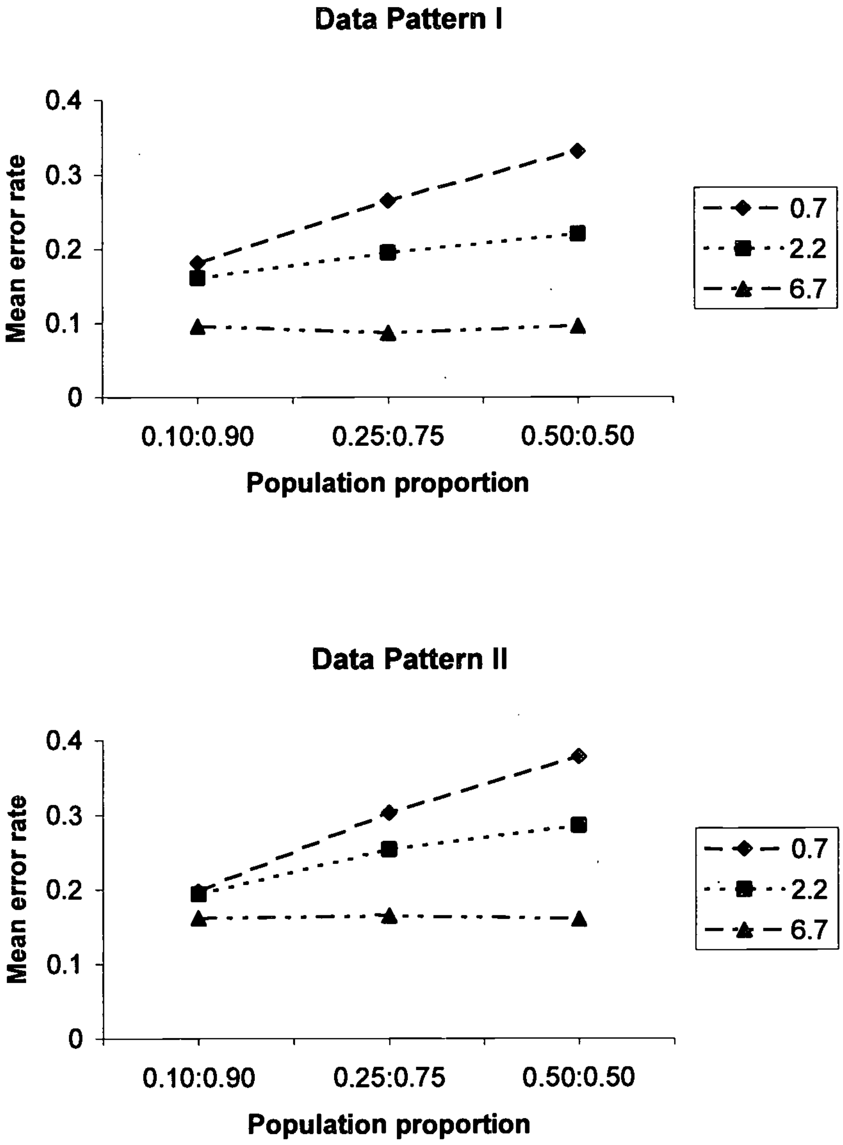
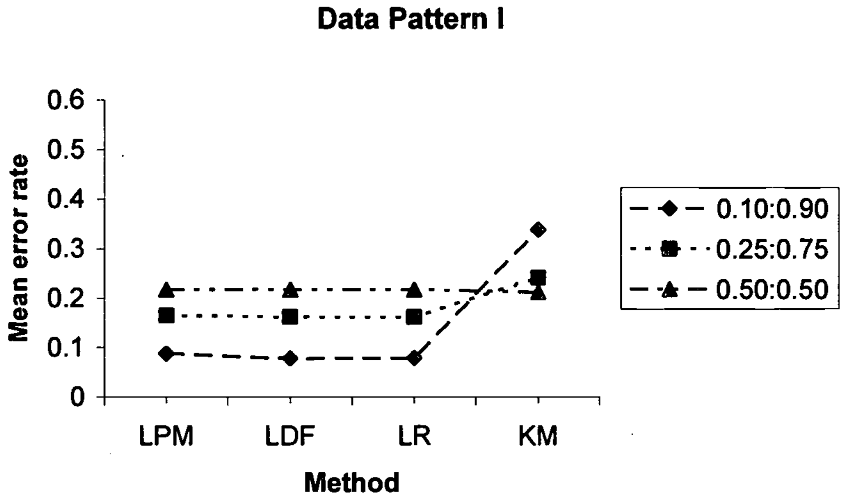


Figure T-2. Method by population proportion interaction on Total error rate.



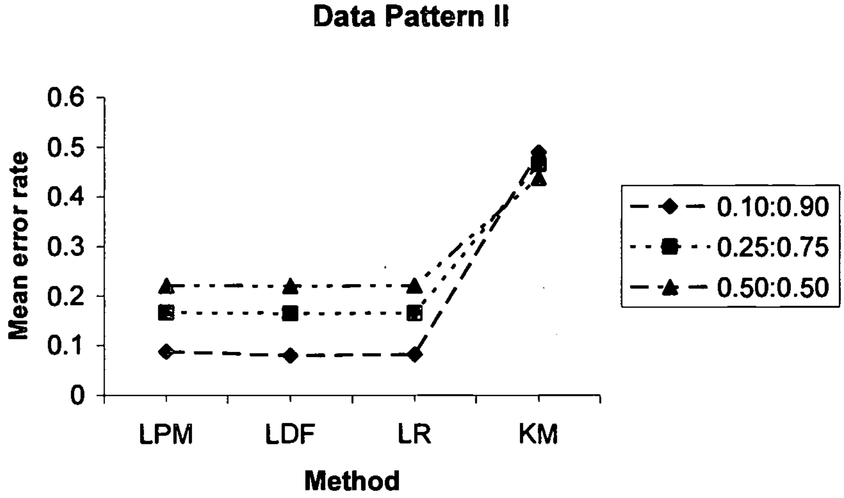
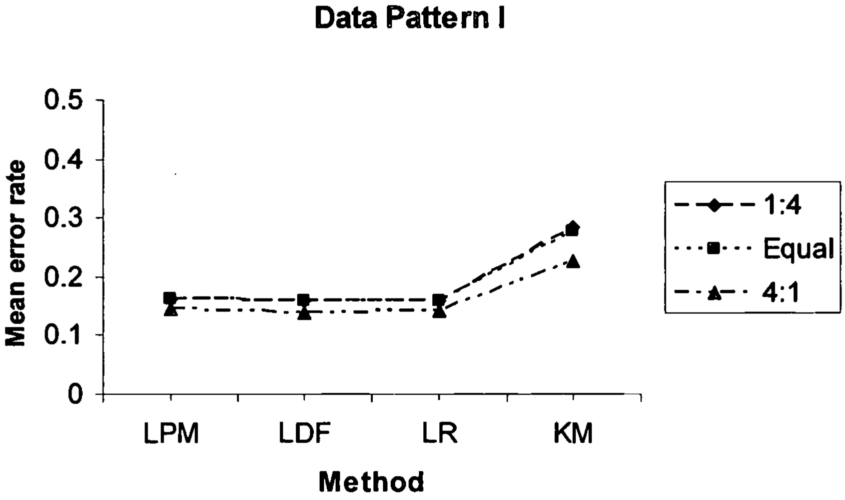


Figure T-3. Method by equality of covariance matrices interaction on Total error rate.



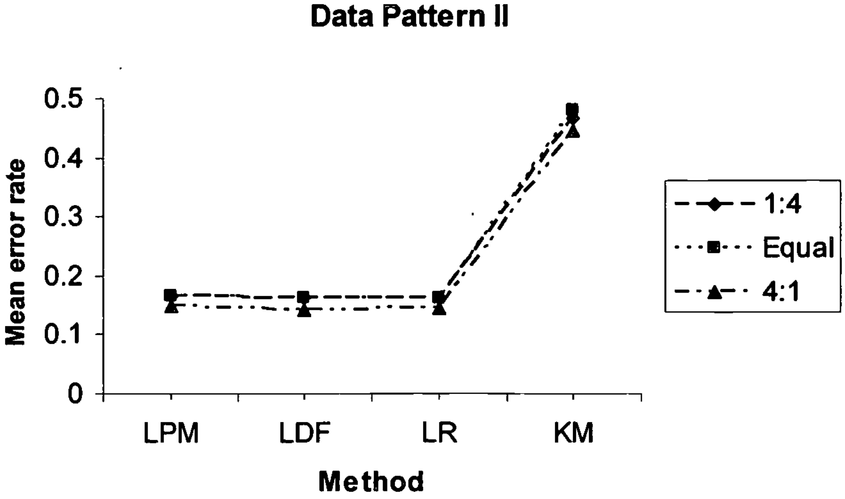


Figure T-4. Method by group separation interaction on Total error rate.

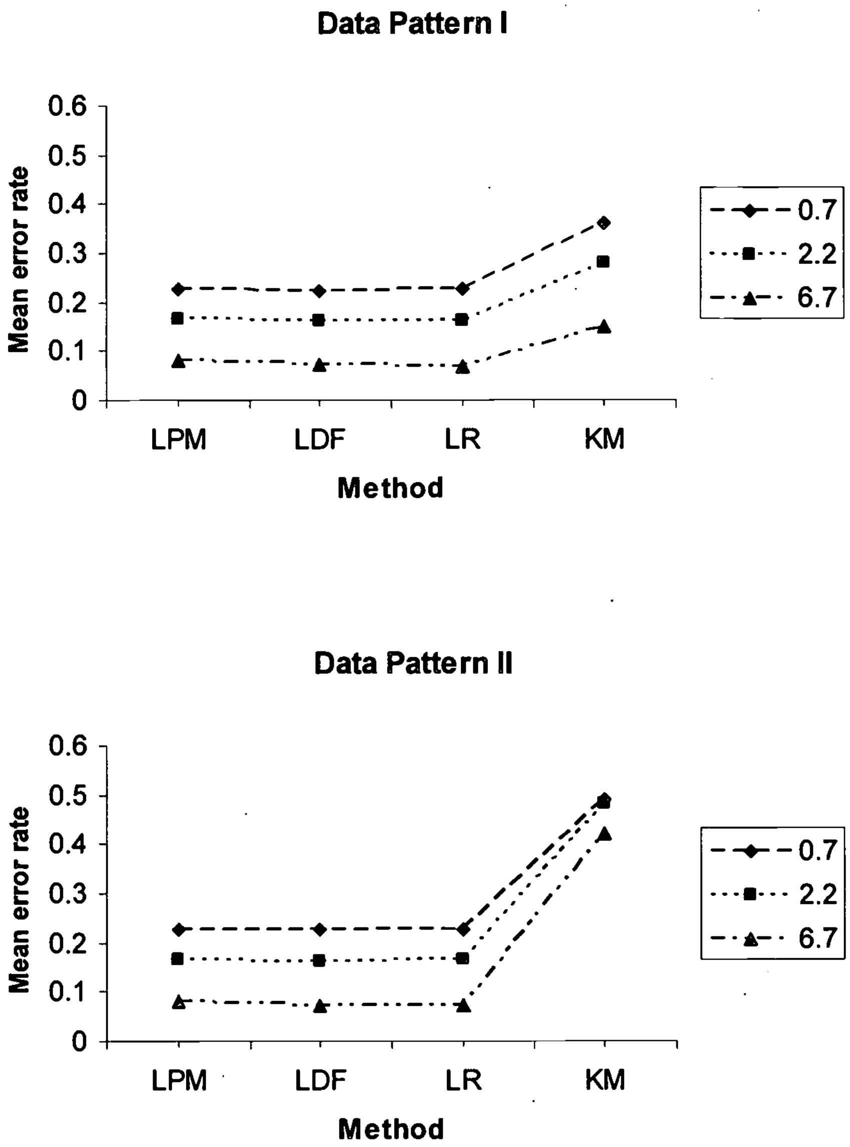
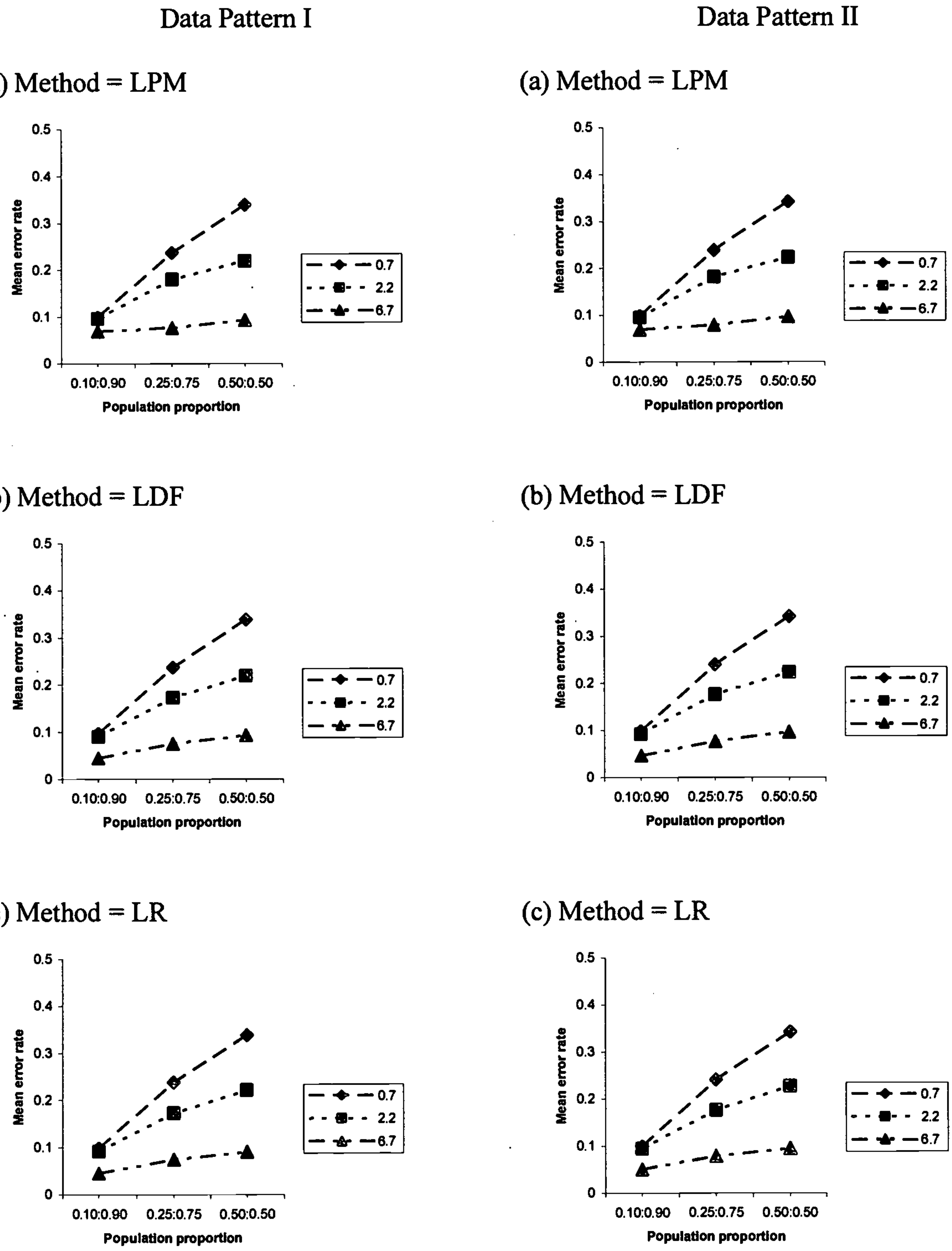


Figure T-5. Method by population proportion by group separation interaction on Total error rate.

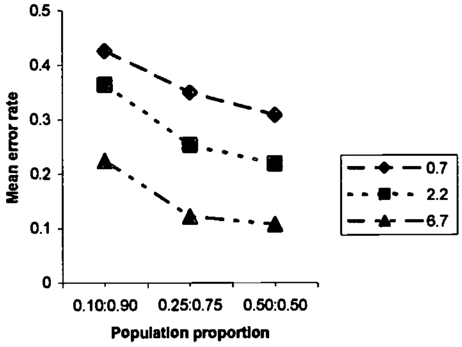


(a) Method = LPM

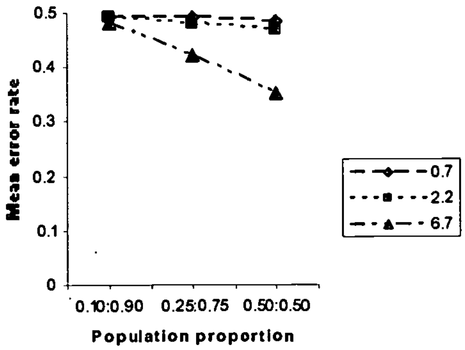
(b) Method = LDF

(c) Method = LR

(d) Method = KM



(d) Method = KM



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034902

I. DOCUMENT IDENTIFICATION:

Title: Comparisons of K-means Clustering with Linear Probability Model, Linear Discriminant Function, and Logistic Regression for Predicting Two-group Membership	
Author(s): Tak-Shing Harry So, Chao-Ying Joanne Peng	
Corporate Source: Paper presented at the 2003 Annual Meeting of American Educational Research Association, Chicago, IL: April 21-25, 2003	Publication Date: 04/25/2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☒	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Tak-Shing Harry So	
Organization/Address: Department of Counseling and Educational Psychology, School of Eduation Indiana University, Bloomington, IN 47405	Telephone: (812) 339-4349	FAX:
	E-Mail Address: tso@indiana.edu	Date: 04/25/2005

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
University of Maryland, College Park
1129 Shriver Lab
College Park, MD 20742