Exploratory Repeated Measures Analysis for Two or More Groups: Review and Update

Stephen W. Looney; William B. Stanley

The American Statistician, Vol. 43, No. 4 (Nov., 1989), 220-225.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198911%2943%3A4%3C220%3AERMAFT%3E2.0.CO%3B2-%23

The American Statistician is currently published by American Statistical Association.

Exploratory Repeated Measures Analysis for Two or More Groups

Review and Update

STEPHEN W. LOONEY and WILLIAM B. STANLEY*

In this article, the exploratory analysis of data from a repeated measures design with one repeated factor and one treatment factor is considered. Recent developments in repeated measures analysis are reviewed and incorporated into an overall strategy for the analysis of such data. An example is given to illustrate the techniques.

KEY WORDS: Adjusted F tests; Interaction; MANOVA; Multiple comparisons; Profile analysis.

1. INTRODUCTION

The repeated measures design is one of the most widely used experimental designs, especially in educational and psychological research. A special type of repeated measures design that is frequently used by researchers is one in which there is one repeated factor (the trials factor) and one treatment factor (the groups factor). Unfortunately, most experimental design textbooks give inadequate coverage to two important aspects of the exploratory analysis of data from such a design: (a) the choice of the appropriate techniques for testing the trial $\times$ group interaction effect and the trial main effect and (b) the appropriate analysis to be used if the trial $\times$ group interaction is found to be significant. [For an exception, see Kirk (1982, chap. 11).] In this article, we address both (a) and (b) and suggest a plan of analysis that incorporates recent developments in repeated measures research. The methodology is illustrated using a set of hypothetical data. The presentation is similar to that of Barcikowski and Robey (1984), who considered repeated measures designs for a single group.

2. REVIEW OF REPEATED MEASURES TECHNIQUES

2.1 Hypotheses To Be Tested

Suppose that observations are obtained on p trials for each of n subjects that are divided into g groups, with n_i subjects in Group i ($1 \leqq i \leqq g$). The hypotheses to be tested in the exploratory analysis of a repeated measures design of this type are usually stated as follows:

H_{OTG}: There is no trial $\times$ group interaction.
H_{OG}: There is no group effect.
H_{OT}: There is no trial effect. (2.1)

Sometimes, the vector of scores for a given subject is referred to as its profile and the analysis of the data from the repeated measures design is referred to as profile analysis (see Harris 1975, pp. 80–84 and 106–108; Morrison 1976, pp. 153–160 and 205–216).

It is important to note the hierarchical nature of the tests of H_{OTG}, H_{OG}, and H_{OT}; namely, H_{OTG} must always be tested first. If H_{OTG} is not rejected, then one proceeds directly to tests of H_{OG} and H_{OT}, followed by the appropriate multiple comparisons. If, on the other hand, H_{OTG} is rejected, then the presence of significant interaction makes it illogical to test either H_{OG} or H_{OT} in the form given in (2.1) (Harris 1975, p. 81). This does not mean that no tests of the group or trial effects can be performed, but that the hypotheses H_{OG} and H_{OT} are not the appropriate hypotheses to test. Alternative techniques for testing the group and trial effects in this case are considered in Section 3.3. The hierarchical nature of H_{OTG}, H_{OG}, and H_{OT} has been ignored by most authors; a notable exception is Rogan, Keselman, and Mendoza (1979, p. 280).

2.2 Hypothesis Testing Procedures

One approach to analyzing the data from the type of repeated measures design considered here is to proceed as in a three-factor mixed model analysis of variance with group and trial being treated as completely crossed fixed factors and subjects being treated as a random factor nested within the group factor (Winer 1971, pp. 518–539). If the trial effects do not satisfy certain "validity conditions" (Huynh and Feldt 1970; Rouanet and Lepine 1970), however, the distributions of the F ratios for testing H_{OTG} and H_{OT} will be distorted, resulting in too frequent rejection of these null hypotheses (Box 1954; Imhof 1962; Huynh and Feldt 1980; Rogan et al. 1979). These validity conditions, which are sometimes referred to as multisample sphericity, can be summarized as follows: (a) the covariance matrices for a suitable set of orthonormalized trial variables must be equal across all levels of the group factor, and (b) the common covariance matrix of these variables must satisfy the sphericity assumption (Huynh 1978, p. 161). In addition to multisample sphericity, the assumption is made that the observations in each group follow a multivariate normal distribution (Scheffe 1959, p. 269).

Several alternatives for testing H_{OTG} and H_{OT} have been

*Stephen W. Looney is Associate Professor, Department of Quantitative Business Analysis, Louisiana State University, Baton Rouge, LA 70803. William B. Stanley is Professor and Chairman, Department of Educational Development, University of Delaware, Newark, DE 19716. This work was partially supported by a summer research grant from the Louisiana State University College of Business Administration. The beneficial comments of six anonymous referees and two associate editors are gratefully acknowledged.

proposed for those cases in which the trial effects do not satisfy the validity conditions. The most commonly used of these are the various adjustments to the univariate approach and the use of multivariate analysis of variance (MANOVA). The adjusted univariate approaches involve using the usual F values for testing H_{OTG} and H_{OT} and modifying the respective degrees of freedom for these test statistics by multiplying them by some estimate of the Geisser–Greenhouse adjustment factor ε (Geisser and Greenhouse 1958, pp. 885–886), which is equal to 1 if the validity conditions are satisfied. In other words, the calculated value of the test statistic F_{cal} is compared with $F[1 - \alpha; \varepsilon' \nu_1, \varepsilon' \nu_2]$, where α is the significance level, ε' is some estimate of ε, and ν_1 and ν_2 are the numerator and denominator degrees of freedom, respectively, for F_{cal}. The F tests for the hypotheses are then performed using the adjusted degrees of freedom. The estimates of ε that will be considered in this article are (a) $e = 1/(p-1)$, the Geisser–Greenhouse (1958, p. 886) conservative estimate; (b) $\hat{\varepsilon}$, the estimate considered by Collier, Baker, Mandeville, and Hayes (1967, p. 343); and (c) $\tilde{\varepsilon}$, the estimate proposed by Huynh and Feldt (1976, p. 75). We also consider the multivariate approach to testing H_{OTG} and H_{OT}; see Greenhouse and Geisser (1959, pp. 105–110) and McCall and Appelbaum (1973, pp. 406–413) for details.

Several authors have compared the various techniques for analyzing repeated measures data. McCall and Appelbaum (1973, p. 414) discussed the relative advantages of the unadjusted, e-adjusted, and $\hat{\varepsilon}$-adjusted univariate tests and the multivariate tests of H_{OTG} and H_{OT} and concluded that the multivariate approach is the best choice when n is large or when the validity conditions are subject to doubt. More recent studies have indicated, however, that certain adjusted univariate procedures can be preferable to the multivariate approach even under these circumstances. For example, Rogan et al. (1979) concluded that (a) for $\varepsilon > .75$, the $\tilde{\varepsilon}$-adjusted procedure is the most powerful test of H_{OTG} and H_{OT}, followed by the $\hat{\varepsilon}$-adjusted and multivariate tests; (b) for $\varepsilon < .75$, the multivariate tests are consistently the most powerful, followed by the $\hat{e}$ and $\bar{e}$-adjusted F tests; and (c) if the parent population is nonnormal [$\chi^2(3)$], then better control of Type I error is achieved by adopting either of the adjusted univariate procedures in favor of the multivariate one. In addition, Maxwell and Arvey (1982) concluded that (a) for $g = 2$, the $\tilde{\varepsilon}$-adjusted F test is preferable to the $\hat{\varepsilon}$-adjusted F; (b) for $g > 2$, the $\hat{\varepsilon}$-adjusted F test tends to be conservative, whereas the $\tilde{\varepsilon}$-adjusted F test tends to be liberal; and (c) either the $\hat{\varepsilon}$- or $\tilde{\varepsilon}$-adjusted F tests can yield reasonable results even when the multivariate test cannot be applied (e.g., when $n - g < p - 1$). These results are consistent with those given by Huynh (1978). The work of all of these authors was taken into account in the analysis procedure recommended in Section 3.

3. RECOMMENDED ANALYSIS PROCEDURES

3.1 Test of No Trial × Group Interaction

The first step in the exploratory analysis of data from a repeated measures design of the type considered here is to test H_{OTG}. Because the univariate and multivariate tests can differ markedly in the type of departures from the null hypothesis that they are able to detect (Davidson 1972; Imhof 1962; Jensen 1982), we suggest that one follow the recommendation of Rouanet and Lepine (1970, p. 161) and perform both a univariate test and a multivariate test of this hypothesis. This approach is further supported by the results of McCall and Applebaum (1973), Romaniuk, Levin, and Hubert (1977), Rogan et al. (1979), and Maxwell and Arvey (1982). We recommend that one allocate half the desired significance level for testing H_{OTG} to each test. In other words, if α denotes the desired significance level, we suggest that one test H_{OTG} at level $\alpha/2$ using the appropriate multivariate test and also at level $\alpha/2$ using the appropriate univariate test. Hypothesis H_{OTG} is then rejected if either test is significant. This approach will guarantee that the overall significance level is no greater than α; however, the power of such a testing procedure has yet to be compared with that of the individual univariate and multivariate tests. Nevertheless, in lieu of a satisfactory criterion for choosing between the two competing approaches, this allocation of significance level seems to be a reasonable compromise.

Several authors (e.g., Huynh and Mandeville 1979; Winer 1971, pp. 595–596) have suggested that univariate testing of repeated measures hypotheses be preceded by tests of the validity conditions to determine whether one should use an adjusted F test instead of the unadjusted test. We do not recommend that one perform these validity tests, however, since it has been shown that they are of no practical use for this purpose (Davidson 1972; Keselman, Rogan, Mendoza, and Breen 1980). Instead, we suggest the following three-step procedure originally proposed by Greenhouse and Geisser (1959), subsequently recommended by Rogan et al. (1979) and Keselman et al. (1980), and modified using the results of Maxwell and Arvey (1982).

The hypothesis H_{OTG} is first tested using the F test with the conservative adjustment, $e = 1/(p-1)$. If the calculated F value is significant using the e-adjusted degrees of freedom, then the result would also be significant using the unadjusted, $\hat{e}$-adjusted, or $\bar{e}$-adjusted degrees of freedom and no further univariate testing is required (Greenhouse and Geisser 1959, p. 110). Similarly, if the F value is not significant using the unadjusted degrees of freedom, then the result also would not be significant using any adjusted degrees of freedom and, again, no further univariate testing is required. If the calculated F value is significant using the unadjusted degrees of freedom, but not significant using the e-adjusted degrees of freedom, then we make the following recommendation based on the results of Rogan et al. (1979) and Maxwell and Arvey (1982): Use the $\tilde{\varepsilon}$-adjusted F test if $g = 2$ or if it is known that $\varepsilon \geqq .75$; otherwise use the more conservative $\hat{\varepsilon}$ adjustment. Thus one should use the $\hat{\varepsilon}$ adjustment if $g > 2$ and nothing is known about ε, as is usually the case (Rogan et al. 1979, pp. 283–284).

As stated earlier, we recommend that one also test H_{OTG} at level $\alpha/2$ using the appropriate multivariate test. In particular, Hotelling's T^2 is used to test H_{OTG} in the case of $g = 2$ groups (Harris 1975, pp. 82–83). If $g > 2$, then one of the general techniques for testing a multivariate linear hypothesis can be used to test H_{OTG}; we suggest that one

use the Pillai–Bartlett trace criterion (Bartlett 1939; Pillai 1955), as recommended by Olsen (1975, p. 583).

As mentioned in Section 2.1, the choice of the appropriate tests for hypotheses H_{OG} and H_{OT} depends on the results of the test of H_{OTG}. We consider this issue in the following two sections.

3.2 No Significant Trial × Group Interaction

If H_{OTG} is not rejected, then an additive model is appropriate and the tests of H_{OT} and H_{OG} can be accomplished by performing tests on the marginal means. For testing the trial effects, the same strategy used in testing H_{OTG} is applied to H_{OT}. The recommendations concerning the choice of a univariate test for testing H_{OTG} also apply when testing H_{OT}; as before, this test is applied at level $\alpha/2$. For the multivariate test, Hotelling's T^2 is used to test H_{OT} for any number of groups (Harris 1975, pp. 107–108). We recommend that this test also be performed at level $\alpha/2$. [Note that this strategy is consistent with that proposed by Barcikowski and Robey (1984, pp. 149–150) for testing the trial effect hypothesis in the case of a single group.] The union–intersection principle (Roy and Bose 1953) can be used to construct simultaneous tests of the contrasts of interest among the marginal trial means, as recommended by Morrison (1976, pp. 134–136) and Boik (1981, p. 254). Alternatively, one can apply the Bonferroni approach using robust tests for marginal trial means based on individual estimates of the contrast variances (Keselman, Rogan, and Games 1981, pp. 166–167), as recommended by Maxwell (1980) and Keselman (1982). It should be noted, however, that the use of this approach may result in seriously inflated Type I error rates, especially under certain combinations of unequal group sizes and unequal covariance matrices across groups (Keselman and Keselman 1988, p. 223).

As far as H_{OG} is concerned, the univariate F test needs no adjustment (Greenhouse and Geisser 1959, p. 101). This hypothesis can be tested using the usual analysis of variance (ANOVA) F test on the means of the sum of the p responses (Harris 1975, pp. 83 and 107), provided that the underlying assumptions of normality and homogeneity of variances across groups appear to be satisfied. If these assumptions are violated, robust alternatives are available; see Brown and Forsythe (1974). An appropriate multiple comparison procedure should be used to test the contrasts of interest when $g > 2$. [See Kirk (1982, pp. 106–126) for details.]

3.3 Significant Trial × Group Interaction

As indicated in Section 2.1, if H_{OTG} is rejected, then it is not appropriate to test either H_{OG} or H_{OT} in the form given in (2.1). One alternative is to attempt to find a transformation of the multivariate data so that the test of H_{OTG} using the transformed data is no longer significant. [See Andrews, Gnanadesikan, and Warner (1971) for an approach to this problem.] If a transformation can be found, then one proceeds as in Section 3.2. If not, we recommend that one follow the suggestion of Morrison (1976, p. 208) and test the hypothesis of equal group means separately for each trial using the usual ANOVA F test. Similarly, the hypothesis of equal trial means can be tested separately within each group using the univariate and multivariate repeated measures techniques described earlier for testing H_{OT}. It is important to note, however, that there are certain circumstances under which it will be known that the validity conditions are trivially satisfied for the trial effects within each group (e.g., when $p = 2$). In this case, one should proceed directly to the unadjusted F test of the trial effect and dispense with the three-step procedure described in Section 3.1.

Once the appropriate testing procedures have been selected, we suggest using a significance level of α/p for tests of the group effect at each level of the trial effect and α/g for tests of the trial effect at each level of the group effect to adequately control the familywise error rates. (For our purposes, a "family" is defined to be the collection of tests associated with a particular factor: there are p such tests for the group factor and g such tests for the trial factor.) Such stringent control of the error rate may lead to extremely low power in detecting small-size treatment effects (Keselman and Keselman 1987), and the investigator should carefully consider the relative costs of Type I and Type II errors before specifying α. [See Kirk (1982, pp. 365–371) for alternative approaches to partitioning the error rate among the tests of the trial effects for each group and among the tests of the group effects for each trial.] Robust multiple comparison methods based on individual estimates of the contrast variances (Keselman et al. 1981, pp. 168–169) can be used to test the contrasts of interest among the group or trial means within each level of the remaining factor.

An alternative to performing a separate test of equal group means for each trial is found by thinking of H_{OG} in terms of the MANOVA hypothesis of equality of g vectors of p-variate means, as follows:

Table 1. Hypothetical Data for Repeated Measures Design

	Group 1			Group 2			Group 3		
Subject	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
1	128	116	206	134	100	152	112	112	244
2	140	108	246	142	124	192	120	104	284
3	108	140	154	162	84	100	140	136	192
4	96	76	94	110	88	40	88	72	132
5	96	88	142	114	96	88	92	84	180
6	88	112	198	122	72	144	100	108	236
7	84	104	74	122	60	20	100	100	112
8	136	100	118	134	124	64	112	96	156
9	116	140	234	166	92	180	144	136	272
10	108	116	174	154	100	120	132	112	212

$$H: \mu_1 = \mu_2 = \cdots = \mu_g,$$

where $\mu_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip}]$ $(1 \leqq i \leqq g)$. Presence of significant interaction indicates that this hypothesis should be rejected. (If H is true, $\mu_{ij} = \mu_{lj}$ for all $1 \leqq i < l \leqq g$, $1 \leqq j \leqq p$, and this would indicate no interaction.) Hence a union–intersection confidence region can be constructed for all differences $\mu_{ij} - \mu_{lj}$ $(1 \leqq i < l \leqq g;\ 1 \leqq j \leqq p)$ to determine which groups are different for each trial (Morrison 1976, pp. 138 and 199). An alternative approach, which might be considerably more powerful, would be to use a Bonferroni region instead of the union–intersection region.

4. EXAMPLE

Consider the hypothetical data given in Table 1 for a repeated measures design in which there are three groups and three trials. Suppose that one is interested in testing the trial × group interaction, trial effect, and group effect. In addition, suppose that if significant trial or group effects are found, it will be of interest to test all possible pairwise comparisons among the levels of the significant factor.

A convenient graphical technique for comparing groups of repeated measurements is to plot the average score on each trial for each group separately and then connect the points. Such a plot is presented in Figure 1 for the data in Table 1. Based on a visual inspection of these graphs, it appears that there are differences among the population profiles of the three groups of subjects, including a significant trial × group interaction.

Since the techniques that we have discussed here are based on the assumption of normality, an attempt was made to verify this assumption. Accordingly, the Shapiro–Wilk (1965) W test was performed for each of the three variates separately, and marginal normality appears to be a reasonable assumption for each one. (The p values are .49, .66, and .92, respectively.) Royston's (1983) procedure for combining these marginal results into a test for multivariate normality also indicated no significant departure ($p \doteq .88$). A similar analysis was performed within each group, and no apparent departures from multivariate normality were detected.

To apply the multivariate test of H_{OTG}, the assumption of equal covariance matrices across the three groups for the successive differences of the original repeated measures must be verified. Box's (1949) modified M criterion indicated no departure from this assumption ($p \doteq .22$), so we may apply the Pillai–Bartlett test to H_{OTG}, yielding $F(4, 54) = 7.36$, $p = .00008$. The conservative e-adjusted F test of H_{OTG} yields $F(2, 27) = 9.90$, $p = .0006$, indicating that any univariate F test of H_{OTG} would be significant at any level of $\alpha \geqq .0006$. Using a per-hypothesis error rate of .05, we see that both the univariate and multivariate tests are significant at level $.05/2 = .025$, and we conclude that there is a significant trial × group interaction. Hence we test for equal group effects separately for each trial and for equal trial effects separately for each group. The results of this analysis for the group effects are as follows: $F(2, 27) = 5.05$, $p = .014$, for Trial 1; $F(2, 27) = 1.71$, $p = .200$, for Trial 2; and $F(2, 27) = 6.34$, $p = .006$, for Trial 3.

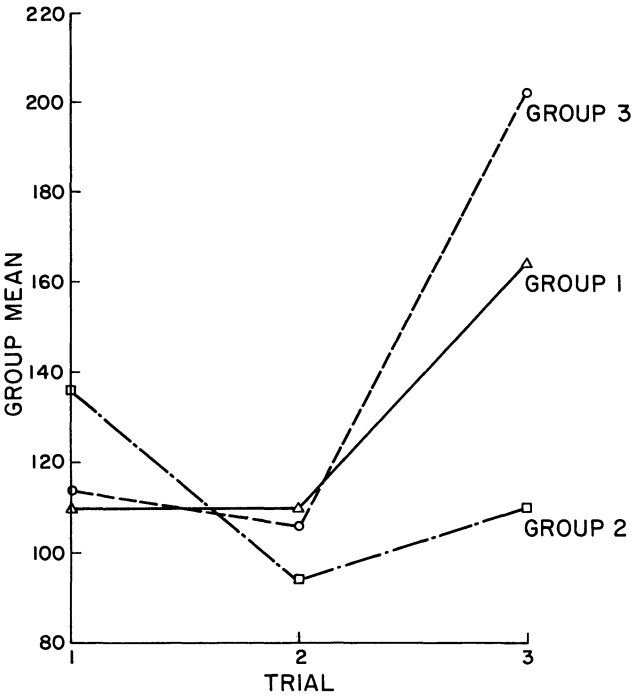

Figure 1. Interaction Plot for Hypothetical Data.

Using a significance level of $\alpha = .05/3 \doteq .017$ for the tests at each trial, we conclude that there are significant differences among the groups for Trials 1 and 3. In Table 2, we present the results for the tests of pairwise comparisons of group means at each trial using the robust method described by Keselman et al. (1981, p. 169). Using a two-tailed per-comparison error rate of $.05/(3 \times 3) \doteq .006$, we find a significant difference only for Groups 2 and 3 at Trial 3 ($p = .002$).

To test for equal trial effects separately for each group, we apply both the univariate and multivariate procedures, using a significance level of $.05/(3 \times 2) \doteq .008$ for each one. For Group 1, the unadjusted F test is significant at this level [$F(2, 18) = 10.29$, $p = .001$], but the e-adjusted F test is not [$F(1, 9) = 10.29$, $p = .011$]. Using the recommendation given in Section 3, since nothing is known about the population value of ε, we rely on the $\hat{\varepsilon}$-adjusted F test ($\hat{\varepsilon} = .69$): $F(1.38, 12.39) = 10.29$, $p = .004$. Therefore, the appropriate adjusted F test of equal trial effects is significant for Group 1 at the .008 level. The Hotelling T^2 test, however, is not significant: $F(2, 8) = 5.48$, $p = .032$. Nevertheless, following the strategy in Section 3, we conclude that there is a significant difference among the trials in Group 1. In Group 2, the unadjusted F test is not significant [$F(2, 18) = 4.48$, $p = .026$], but the Hotelling T^2 test is [$F(2, 8) = 13.22$, $p = .003$]. In accordance with the strategy outlined in Section 3, we conclude that there is also a significant trial effect in Group 2.

Table 2. Two-Tailed p Values for Pairwise Comparisons of Group Means at Each Trial

	Trial		
Group comparison	1	2	3
1 versus 2	.009	.093	.052
1 versus 3	.655	.663	.161
2 versus 3	.023	.200	.002

Table 3. Two-Tailed p Values for Pairwise Comparisons of Trial Means in Each Group

	Group		
Trial comparison	1	2	3
1 versus 2	1.000	.001	.021
1 versus 3	.008	.136	<.001
2 versus 3	.007	.370	<.001

For Group 3, both the e-adjusted F test [$F(1, 9) = 33.60$, $p = .0003$] and the Hotelling T^2 test [$F(2, 8) = 19.43$, $p = .0008$] are significant at the .008 level. Thus a significant difference among the trials is also indicated for Group 3. In Table 3, we present the results for the tests of pairwise comparisons of trial means in each group using the robust method described by Keselman et al. (1981, p. 169). Using a two-tailed per-comparison error rate of $.05/(3 \times 3) \doteq .006$, we find that Trials 1 and 2 are significantly different in Group 2 and that Trial 3 is significantly different from Trials 1 and 2 in Group 3.

In summary, our analysis leads to the following conclusions: (a) there is a significant trial × group interaction, (b) Groups 2 and 3 are significantly different at Trial 3, (c) Trials 1 and 2 are significantly different in Group 2, and (d) Trial 3 is significantly different from Trials 1 and 2 in Group 3. No other significant differences could be found.

5. SUMMARY

In this article we have reviewed recent developments in repeated measures methodology and incorporated these into a framework for the exploratory analysis of data from repeated measures designs with a single repeated factor (trial) and a single nonrepeated factor (group). These developments include new advances in adjusting the univariate F tests for these designs, as well as results on the relative merits of the univariate approaches versus the multivariate approach. We suggest that one use appropriate univariate tests as well as multivariate tests, since the sensitivities of the univariate and multivariate approaches are not related in any consistent way. We have also made recommendations concerning the appropriate analysis to be used in the presence of significant trial × group interaction, and we illustrated our proposed strategy with an example.

[*Received August 1986. Revised March 1989.*]

REFERENCES

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971), "Transformations of Multivariate Data," *Biometrics*, 27, 825–840.

Barcikowski, R. S., and Robey, R. R. (1984), "Decisions in Single Group Repeated Measures Analysis: Statistical Tests and Three Computer Packages," *The American Statistician*, 38, 148–150.

Bartlett, M. S. (1939), "A Note on Tests of Significance in Multivariate Analysis," *Proceedings of the Cambridge Philosophical Society*, 35, 180–185.

Boik, R. J. (1981), "A Priori Tests in Repeated Measures Designs: Effects of Nonsphericity," *Psychometrika*, 46, 241–255.

Box, G. E. P. (1949), "A General Distribution Theory for a Class of Likelihood Criteria," *Biometrika*, 36, 317–346.

——— (1954), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II: Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification," *The Annals of Mathematical Statistics*, 25, 484–498.

Brown, M. B., and Forsythe, A. B. (1974), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics*, 16, 129–132.

Collier, R. O., Jr., Baker, F. B., Mandeville, G. K., and Hayes, T. F. (1967), "Estimates of Test Size for Several Test Procedures Based on Conventional Variance Ratios in the Repeated Measures Design," *Psychometrika*, 32, 339–353.

Davidson, M. L. (1972), "Univariate Versus Multivariate Tests in Repeated Measures Experiments," *Psychological Bulletin*, 77, 446–452.

Geisser, S., and Greenhouse, S. W. (1958), "An Extension of Box's Results on the Use of the F Distribution in Multivariate Anaysis," *The Annals of Mathematical Statistics*, 29, 885–891.

Greenhouse, S. W., and Geisser, S. (1959), "On Methods in the Analysis of Profile Data," *Psychometrika*, 24, 95–112.

Harris, R. J. (1975), *A Primer of Multivariate Statistics*, New York: Academic Press.

Huynh, H. (1978), "Some Approximate Tests for Repeated Measurement Designs," *Psychometrika*, 43, 161–175.

Huynh, H., and Feldt, L. S. (1970), "Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions," *Journal of the American Statistical Association*, 65, 1582–1589.

——— (1976), "Estimation of the Box Correction for Degrees of Freedom From Sample Data in the Randomized Block and Split-Plot Designs," *Journal of Educational Statistics*, 1, 69–82.

——— (1980), "Performance of Traditional F Tests in Repeated Measures Designs Under Covariance Heterogeneity," *Communications in Statistics, Part A—Theory and Methods*, 9, 61–74.

Huynh, H., and Mandeville, G. K. (1979), "Validity Conditions in Repeated-Measures Designs," *Psychological Bulletin*, 86, 964–973.

Imhof, J. P. (1962), "Testing the Hypothesis of No Fixed Main-Effects in Scheffe's Mixed Model," *The Annals of Mathematical Statistics*, 33, 1085–1095.

Jensen, D. R. (1982), "Efficiency and Robustness in the Use of Repeated Measurements," *Biometrics*, 38, 813–825.

Keselman, H. J. (1982), "Multiple Comparisons for Repeated Measures Means," *Multivariate Behavioral Research*, 17, 87–92.

Keselman, H. J., and Keselman, J. C. (1987), "Type I Error Control and the Power to Detect Factorial Effects," *British Journal of Mathematical and Statistical Psychology*, 40, 196–208.

——— (1988), "Repeated Measures Multiple Comparison Procedures: Effects of Violating Multisample Sphericity in Unbalanced Designs," *Journal of Educational Statistics*, 13, 215–226.

Keselman, H. J., Rogan, J. C., and Games, P. A. (1981), "Robust Tests of Repeated Measures Means in Educational and Psychological Research," *Educational and Psychological Measurement*, 41, 163–173.

Keselman, H. J., Rogan, J. C., Mendoza, J. L., and Breen, L. J. (1980), "Testing the Validity Conditions of Repeated Measures F Tests," *Psychological Bulletin*, 87, 479–481.

Kirk, R. E. (1982), *Experimental Design* (2nd ed.), Belmont, CA: Brooks/Cole.

Maxwell, S. E. (1980), "Pairwise Multiple Comparisons in Repeated Measures Designs," *Journal of Educational Statistics*, 5, 269–287.

Maxwell, S. E., and Arvey, R. D. (1982), "Small Sample Profile Analysis With Many Variables," *Psychological Bulletin*, 92, 778–785.

McCall, R. B., and Appelbaum, M. I. (1973), "Bias in the Analysis of Repeated-Measures Designs: Some Alternative Approaches," *Child Development*, 44, 401–415.

Morrison, D. F. (1976), *Multivariate Statistical Methods* (2nd ed.), New York: McGraw-Hill.

Olsen, C. L. (1976), "On Choosing a Test Statistic in Multivariate Analysis of Variance," *Psychological Bulletin*, 83, 579–586.

Pillai, K. C. S. (1955), "Some New Test Criteria in Multivariate Analysis," *The Annals of Mathematical Statistics*, 26, 117–121.

Rogan, J. C., Keselman, H. J., and Mendoza, J. L. (1979), "Analysis of Repeated Measurements," *British Journal of Mathematical and Statistical Psychology*, 32, 269–286.

Romaniuk, J. G., Levin, J. R., and Hubert, L. J. (1977), "Hypothesis-Testing Procedures in Repeated-Measures Designs: On the Road Map Not Taken," *Child Development*, 48, 1757–1760.

Rouanet, H., and Lepine, D. (1970), "Comparison Between Treatments in a Repeated Measures Design: ANOVA and Multivariate Methods," *British Journal of Mathematical and Statistical Psychology*, 23, 147–163.

Roy, S. N., and Bose, R. C. (1953), "Simultaneous Confidence Interval Estimation," *The Annals of Mathematical Statistics*, 24, 513–536.

Royston, J. P. (1983), "Some Techniques for Assessing Multivariate Normality Based on the Shapiro–Wilk *W*," *Applied Statistics*, 32, 121–133.

Scheffe, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Shapiro, S. S., and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611.

Winer, B. J. (1971), *Statistical Principles in Experimental Design* (2nd ed.), New York: McGraw-Hill.