# Statistical Practises of Educational Researchers:

# An Analysis of Their ANOVA, MANOVA and ANCOVA Analyses

by

H.J. Keselman[1]

University of Manitoba

Carl J Huberty
University of Georgia

Lisa M. Lix
Private Scholar

Stephen Olejnik
University of Georgia

Robert A. Cribbie
University of Manitoba

Barbara Donahue
University of Georgia

Rhonda K. Kowalchuk
University of Manitoba

Laureen L. Lowman
University of Georgia

Martha D. Petoskey
University of Georgia

and

Joanne C. Keselman
University of Manitoba

Joel R. Levin
University of Wisconsin, Madison

Abstract

Articles published in several prominent educational journals were examined to investigate the use of data-analysis tools by researchers in four research paradigms: between-subjects univariate designs, between-subjects multivariate designs, repeated measures designs, and covariance designs. In addition to examining specific details pertaining to the research design (e.g., sample size, group size equality/inequality) and methods employed for data analysis, we also catalogued whether: (a) validity assumptions were examined, (b) effect size indices were reported, (c) sample sizes were selected based on power considerations, and (d) appropriate textbooks and/or articles were cited to communicate the nature of the analyses that were performed. Our analyses imply that researchers rarely verify that validity assumptions are satisfied and accordingly typically use analyses that are nonrobust to assumption violations. In addition, researchers rarely report effect size statistics, nor do they routinely perform power analyses to determine sample size requirements. We offer many recommendations to rectify these shortcomings.

**Statistical Practises of Educational Researchers:**

**An Analysis of Their ANOVA, MANOVA and ANCOVA Analyses**

It is well known that the volume of published educational research is increasing at a very rapid pace. As a consequence of the expansion of the field, qualitative and quantitative reviews of the literature are becoming more common. These reviews typically focus on summarizing the results of research in particular areas of scientific inquiry (e.g., academic achievement or English as a second language) as a means of highlighting important findings and identifying gaps in the literature. Less common, but equally important, are reviews that focus on the research process, that is, the methods by which a research topic is addressed, including research design and statistical analysis issues.

Methodological research reviews have a long history (e.g., Edgington, 1964; Elmore & Woehlke, 1988, 1998; Goodwin & Goodwin, 1985a, 1985b; West, Carmody, & Stallings, 1983). One purpose of these reviews has been the identification of trends in data-analytic practice. The documentation of such trends has a two-fold purpose: (a) it can form the basis for recommending improvements in research practice, and (b) it can be used as a guide for the types of inferential procedures that should be taught in methodological courses, so that students have adequate skills to interpret the published literature of a discipline and to carry out their own projects.

One consistent finding of methodological research reviews is that a substantial gap often exists between the inferential methods that are recommended in the statistical research literature and those techniques that are actually adopted by applied researchers (Goodwin & Goodwin, 1985b; Ridgeway, Dunston, & Qian, 1993). The practice of relying on traditional methods of analysis is, however, dangerous. The field of statistics is by no means static; improvements in statistical procedures occur on a regular basis. In particular, applied statisticians have devoted a great deal of effort to understanding the operating characteristics of statistical procedures when the distributional assumptions that underlie a particular procedure are not likely to be satisfied. It is common knowledge that under certain data-analytic conditions, statistical procedures will not produce valid results. The applied researcher who routinely adopts a traditional procedure without

giving thought to its associated assumptions may unwittingly be filling the literature with nonreplicable results.

Every inferential statistical tool is founded on a set of core assumptions. As long as the assumptions are satisfied, the tool will function as intended. When the assumptions are violated, however, the tool may mislead. It is well known that the general class of analysis of variance (ANOVA) tools frequently applied by educational researchers, and considered in this article, includes at least three key distributional assumptions. For all cases the outcome measure $Y_{ki}$ (or "score") associated with the i-th individual within the k-th group is normally and independently distributed, with an mean of $\mu_k$ and a variance of $\sigma^2$. Importantly, because $\sigma^2$ does not include a k subscript, this indicates that the score variances within all groups are equal (variance homogeneity).

Only if these three assumptions are met can the traditional F tests of mean differences be validly interpreted, for without the assumptions (or barring strong evidence that adequate compensation for them has been made), it can be -- and has been -- shown that the resulting "significance" probabilities (p-values) are, at best, somewhat different from what they should be and, at worst, worthless. Concretely, what this means is that an assumptions-violated test of group effects might yield a F ratio with a corresponding significance probability of $p = .04$, which (based on an a priori Type I error probability of .05) would lead a researcher to conclude that there are statistically nonchance differences among the K groups. However, and unknown to the unsuspecting researcher, the "true" probability of the obtained results, given a no-difference hypothesis and violated assumptions, could perhaps be $p = .37$, contrarily suggesting that the observed differences are likely due to chance. And, of course, the converse is also true: A significance probability that leads a researcher to a no-difference conclusion might actually be a case of an inflated Type II error probability stemming from the violated distributional assumptions.

The "bottom line" here is that in situations where a standard parametric statistical test's assumptions are suspect, conducting the test anyway can be a highly dangerous practice. In this

article, we not only remind the reader of the potential for this danger but, in addition, provide evidence that the vast majority of educational researchers are conducting their statistical analyses without taking into account the distributional assumptions of the procedures they are using.

Thus one purpose of the following content analyses (based on a sampling of published empirical studies) was to describe the practices of educational researchers with respect to inferential analyses in popular research paradigms. The literatures reviewed encompass designs that are commonly used by educational researchers -- that is, univariate and multivariate independent (between-subjects) and correlated groups (within-subjects) designs that may contain covariates. In addition to providing information on the use of statistical procedures, the content analyses focused on topics that are of current concern to applied researchers, such as power analysis techniques and problems of assumption violations. Furthermore, consideration was given to the methodological sources that applied researchers use, by examining references to specific statistical citations. Our second purpose, based on the findings of our reviews, is to present recommendations for reporting research results and for obtaining valid methods of analysis.

Prominent educational and behavioral science research journals were selected for review.[1] An enumeration of the journals reviewed can be found in Table 1. These journals were chosen because they publish empirical research, are highly regarded within the fields of education and psychology, and represent different education subdisciplines. To the extent possible, all of the articles published in the 1995/1994 issue of each journal were reviewed by the authors.

## The Analysis of Between-Subjects Univariate Designs

Past research has shown that the ANOVA F test is the most popular data-analytic technique among educational researchers (Elmore & Woehlke, 1998; Goodwin & Goodwin, 1985a) and that it is used most frequently within the context of one-way and factorial between-subjects univariate designs. However, researchers should be aware that although the ANOVA F test is the conventional approach for conducting tests of mean equality in between-subjects designs, it is not necessarily a valid approach, due to its reliance on the assumptions of normality and variance homogeneity. Specifically, recent surveys indicate that the data collected by

educational and psychological researchers rarely if ever come from populations that are characterized by the normal density function or by homogeneous variances (Micceri, 1989; Wilcox, Charlin & Thompson, 1986). Hence, as previously indicated, the validity of statistical procedures that assume this underlying structure to the data is seriously in question. Specifically, the effect of using ANOVA when the data are nonnormal and/or heterogeneous is a distortion in the rates of Type I and/or Type II errors (or, the power of the test), particularly when group sizes are unequal.

In this content analysis we examined the method(s) adopted for testing hypotheses of mean equality involving main, interaction, and/or simple between-subjects effects. Methods for testing omnibus (overall) hypotheses could include the ANOVA F test or an alternative to the F test. Alternative test procedures could include the nonparametric Kruskal-Wallis test (Kruskal & Wallis, 1952) or the Mann-Whitney U test (in the case of two groups), as well as various parametric procedures such as the Brown and Forsythe, James, and Welch tests (see Coombs, Algina & Oltman, 1996) which are all relatively insensitive to the presence of variance heterogeneity. Trend analysis may also be used in cases where the levels of the between-subjects factor(s) are quantitative, rather than qualitative in nature. As well, planned (a priori) contrasts on the data may be used to answer very specific research questions concerning one's data.

The use of multiple comparison procedures (MCPs) for testing hypotheses concerning pairs of between-subjects means was also examined. The specific strategy adopted to control either the familywise rate of error (FWE) or the per-comparison rate of error (PCE) was identified, as was the type of test statistic used. In between-subjects designs, the pairwise comparison test statistic may be computed in different ways, depending on the assumptions the researcher is willing to make about the data (see Maxwell & Delaney, 1990, pp. 144-150). For example, in a one-way design, one test statistic (which we will call single-error) incorporates the error term from the omnibus test of the between-subjects effect. Accordingly, the variance homogeneity assumption must be satisfied for such an approach to provide valid tests of pairwise comparisons. The alternative (separate-error) uses an error term based on only data associated

with the particular levels of the between-subjects factor that are being compared. In the latter approach, which does not assume homogeneity across all factor levels, each pairwise comparison statistic has a separate-error term.

In unbalanced (unequal cell sizes) factorial designs, also known as nonorthogonal designs, the sums of squares (SS) for marginal (e.g., main) effects may be computed in different ways. That is, tests of weighted or unweighted means may be performed depending on the hypotheses of interest to the researcher (see Carlson & Timm, 1974).

Research Design Features and Methods of Analysis

Table 2 contains information pertaining to design characteristics of the 61 between-subjects articles which were examined in this content analysis. One-way designs (59.0%) were more popular than factorial designs (47.5%). However, it should be noted that there was some overlap with respect to this classification, as four articles reported the use of both types of designs.

Overall, unbalanced designs were more common than balanced designs. This is particularly evident in the case of studies involving factorial designs, where almost three-quarters of those identified (72.4%) were comprised of cells containing unequal numbers of units of analysis. Of the 23 one-way studies in which an unbalanced design was used, the ratio of the largest to the smallest group size was greater than 3 in 43.5% of these. Of the 21 unbalanced factorial studies, the ratio of the largest to the smallest cell size was greater than 3 in 38.1% of these.

Table 2 also contains information pertaining to the methods of inferential analysis in the studies which incorporated a between-subjects univariate design. The ANOVA F test was overwhelmingly favored, and was used by researchers in more than 90% of the articles. A nonparametric analysis was performed by the authors of only four articles; in each of these, one-way designs were under investigation. Planned contrasts were reported in two articles, in both cases for assessing an effect in a one-way design. Trend analysis was used by the authors of one article, also in relation to the analysis of a one-way design.

In only 3 of the 21 articles in which a nonorthogonal design was used did the authors report the method adopted to compute the SS for marginal effects. In two of these, unweighted means were adopted and in one weighted means were used.

In two articles, both involving one-way designs, the authors did not conduct a test of the omnibus hypothesis, and instead proceeded directly to pairwise mean comparisons. In total, 29 articles reported the use of a MCP (46.8%). Tukey's procedure was most popular (27.6%), followed by the Newman-Keuls method (20.7%) (see Kirk, 1995 for MCP references). In only three instances (10.3%) did the author(s) conduct unprotected multiple t-tests, which allow for control of the PCE rather than the FWE.

Little difference existed in the popularity of MCPs for the analysis of one-way and factorial designs; in both cases Tukey's procedure was favored. However, Duncan's procedure was only used for testing hypotheses involving pairs of means in one-way designs and the Newman-Keuls procedure was more popular in factorial designs than in one-way designs.

It has been shown that both the Fisher and Newman-Keuls procedures cannot control the FWR when more than three means are compared in a pairwise fashion (Keselman, Keselman, & Games, 1991; Levin, Serlin, & Seaman, 1994). Despite this, half of the studies in which the Newman-Keuls procedure was adopted contained more than three means, while Fisher's procedure was used in one such study.

MCPs were used in factorial designs more often to test for differences in pairs of marginal means ($n = 9$), than to test pairs of simple means ($n = 6$). Finally, with respect to the test statistic used in the MCP analyses, in only one article was it possible to discern that a separate-error test statistic had been adopted. In this case, which involved a one-way design, multiple t tests were conducted, and the authors did not perform a preliminary omnibus analysis.

Assessment of Validity Assumptions

With respect to the assessment of validity assumptions, our first task was to examine possible departures from variance heterogeneity. Thirteen of the 61 articles which incorporated between-subjects univariate designs did not report group or cell standard deviations for any of the

dependent variables under investigation. For the remaining articles, we focused our attention on at most the first five variables that were subjected to analysis in order to limit the data set to a manageable size. For one-way designs, we collected standard deviation information for 86 dependent variables. The average value of the ratio of the largest to smallest standard deviation was 2.0 (SD = 2.6), with a median of 1.5. Several extreme ratio values were noted in the one-way designs, with a maximum ratio of 23.8. In the factorial studies, information was obtained for 85 dependent variables, with a mean ratio of 2.8 (SD = 4.2), a median of 1.7, and a maximum ratio of 29.4.

For one-way designs, a positive relationship between group sizes and standard deviations existed for 31.3% of the dependent variables, a negative relationship was identified for 22.1%, no discernible pattern was observed for 15.1%, and this classification was not applicable for 25.6% of the dependent variables because group sizes were equal. For five dependent variables it was not possible to categorize this relationship because group size information was not provided. For factorial designs, a negative relationship between cell sizes and standard deviations was revealed for 23.5% of the dependent variables, a positive relationship was evident for 14.1%, and no relationship was evident for 31.8% of the dependent variables. As well, this relationship was not applicable for 14.1% of the dependent variables because the design was balanced.

In 12 articles (19.7%), the author(s) indicated some concern for distributional assumption violations. Normality was a consideration in seven articles, although no specific tests for violations of this assumption were reported; rather, it appears that normality was assessed by descriptive measures only. Variance homogeneity was evaluated in five articles, and it was specifically stated that this assumption was tested in three of these articles. Only one article considered both assumptions simultaneously.

The authors of these articles used a variety of methods to deal with assumption violations. In total, five studies relied on transformations; typically, these were used where the dependent variables of interest were measured using a percentage scale. A nonparametric procedure was adopted in two articles, in one because the dependent variable under investigation was skewed,

and in the other because variances were heterogeneous. One set of authors tested for heterogeneity using Levene's (1960) test and obtained a significant result, but chose to proceed with use of the ANOVA F test. In two articles where skewness was due to outliers, these values were Winsorized; that is, the extreme scores were replaced with less extreme values. In one case, the authors chose to redesign the study in order to avoid dealing with nonnormal data. Thus, although a $2 \times 4$ factorial design was originally employed, it was reduced to a $2 \times 2$ design because of nonnormality due to floor effects in four cells of the design. Finally, the authors of one study elected to convert a continuous dependent variable to a categorical variable, and then they conducted a frequency analysis rather than a means analysis due to the existence of skewness in the data.

Power/Effect Size Analysis

The issue of power and/or effect size calculations arose in only 10 articles (16.1%). Effect sizes were calculated in six of these, but the statistic used was not routinely reported and main effects were more often of interest than interactions. The authors of two articles were concerned that the power to detect an interaction might be low, and thus performed post hoc analyses of power. The authors of one article reported that although the independent variable under investigation was quantitative in nature, it was converted to a categorical variable and the ANOVA F test was used instead of regression analysis. This was done because the authors felt that the former approach would result in greater statistical power than the latter; however, no empirical support for this premise was given.

Software Packages/Statistical Citations

The statistical software package used in data analysis was specified in only five articles. In three of these, SPSS (Norusis, 1993) was used while SYSTAT (Wilkinson, 1988) and SAS (SAS Institute, 1990) were each used once. A variety of statistical sources were cited in the articles. However, no single source was used with great frequency and thus this component of the analysis was unrevealing.

Conclusions and Recommendations Concerning Between-Subjects Univariate Designs

This review reveals that behavioral science researchers use between-subjects univariate designs in a variety of contexts. Investigations involving a single between-subjects factor were favored slightly more than those in which the effects of multiple factors were jointly considered, although in both cases, designs with unequal group sizes were more popular than designs with equal group sizes.

As anticipated, the ANOVA F test was the method of choice for examining group effects, despite its reliance on the stringent assumptions of normality and variance homogeneity. This is a disturbing trend, as Lix, Keselman, and Keselman (1996), in a quantitative review of the effects of assumption violations on the ANOVA F test in one-way designs, found very few instances in which this conventional method of analysis was appropriate. Although the ANOVA F test may be relatively insensitive to violations of the normality assumption in terms of Type I error control, it is highly sensitive to differences in population variances. This sensitivity is accentuated when group sizes are unequal. Similar findings have been reported by Keselman, Carriere, and Lix (1995) and Milligan, Wong, and Thompson (1987) with respect to factorial designs, regardless of the method used to compute the sums of squares for marginal effects. Normality does, however, have important implications for the control of Type II errors (Wilcox, 1995).

The routine use of the F test in the face of assumption violations may stem from the fact that behavioral science researchers do not appear to give a great deal of thought to assumption violations, as less than 20% of the articles considered in this review made mention of this issue. When it was clear that assumptions were considered, normality was more likely to be of concern than variance homogeneity, and transformations were typically used as a means of normalizing the distribution of responses. Although the adoption of a nonparametric procedure may be useful when the normality assumption is untenable, it is not good practice when the assumption of variance homogeneity is suspect. The Lix et al. (1996) review showed that the Kruskal-Wallis test (Kruskal & Wallis, 1952) is highly sensitive to unequal variances.

It is equally important to consider the underlying distributional assumptions when pairwise comparisons of means or other contrasts are performed on the data. In only one paper was the choice of a test statistic specified (in that case a separate error statistic was used), and thus it was difficult to determine what assumptions the majority of researchers were making about the data in testing hypotheses involving pairs of means.

It is interesting to note that in only two studies did the authors not elect to perform omnibus tests of between-subjects effects. Rather, the more common practice was to perform one or more omnibus tests, which, if significant, were followed by simple effect tests and/or pairwise comparisons of means.

As anticipated, effect sizes were almost never reported along with p-values, despite encouragement to do so by the most recent edition of the American Psychological Association's (1994) Publication Manual. Moreover, indications of the magnitude of interaction effects were extremely rare. Finally, it should be noted that in all instances where effect sizes were given, a statistically significant result was obtained.

We feel there are a number of ways in which behavioral science researchers can improve their analyses of between-subjects univariate designs. We strongly encourage: (a) selecting robust methods for conducting omnibus tests and contrasts, (b) conducting focused tests of hypotheses, and (c) routinely reporting measures of effect.

With respect to the first point, many studies have demonstrated that the ANOVA F test is very frequently inappropriate to test for the presence of group mean differences in between-subjects designs (see e.g., Wilcox, 1987). Despite these repeated cautionary notes, behavioral science researchers have clearly not taken this message to heart. It is strongly recommended that test procedures that have been designed specifically for use in the presence of variance heterogeneity and/or nonnormality be adopted on a routine basis. A number of research reviews give clear information on selection of robust methods. A good starting point is the paper by Lix et al. (1996), which documents the deficiencies of the F test (see also Harwell, Rubenstein, Hayes, & Olds, 1992) and provides clear guidelines on the conditions under which various robust

procedures -- including the Welch and James procedures -- will exhibit optimal results. Also included in that paper is a discussion of computer programs that will perform these tests. Procedures that are robust to both variance heterogeneity and nonnormality are considered by Lix and Keselman (1998). A discussion of robust methods for use in factorial designs can be found in Keselman et al. (1995) and Keselman, Kowalchuk, and Lix (1998) -- see also Hsiung and Olejnik (1994b). The application of robust methods for conducting pairwise mean comparisons is considered by Keselman, Lix, and Kowalchuk (1997), Lix and Keselman (1995), and Olejnik and Hess(1997).

With respect to the second point, behavioral science researchers need to critically evaluate the usefulness of conducting preliminary omnibus tests of main and/or interaction effects. As Olejnik and Huberty (1993) note, "the most important limitation of the omnibus F-test is that it is so general that it typically does not address an interesting substantive question" (p. 7). It was typically the case that if a significant omnibus result was obtained, it was followed with additional tests to provide further information on the nature of the effect, such as pairwise mean comparisons. It is entirely possible to bypass the omnibus test and proceed directly to simple effect tests or pairwise comparisons, although a few MCPs do incorporate a preliminary test. A comprehensive discussion of the use of planned contrasts for data analysis can be found in most popular research methods/statistics textbooks, including Kirk (1995), and Maxwell and Delaney (1990), as well as in the work of Hsiung and Olejnik (1994a).

With respect to the third point, numerous sources have discussed the need for reporting a measure of effect size along with a p-value, in order to allow the reader to distinguish between those results that are "practically" significant and those that are only "statistically" significant. Although it is encouraging that a small number of the articles reviewed in the current content analysis reported a measure of effect or some form of power analysis, this type of information needs to be routinely reported. Educational researchers have at their disposal numerous sources on this topic, including Cohen (1992), Kirk (1996), and O'Brien and Muller (1993), as well as the recent compendium by Harlow, Muliak, and Steiger (1997).

## The Analysis of Between-Subjects Multivariate Designs

Univariate ANOVA actually involves more than one characteristic of the (experimental) units involved. There is one outcome variable; but there can be more than one grouping variable. It is the effect of the grouping variable(s) on the outcome variable that is of interest to the researcher who employs ANOVA techniques. Multivariate analysis of variance (MANOVA) can have one or more grouping variables, but would include multiple outcome variables (say, P in number). It is the effect of the grouping variable(s) on the collection of outcome variables that is of interest to the researcher who uses MANOVA techniques. Just as in the case of an ANOVA with one grouping variable, the interest in a MANOVA with one grouping variable is group comparison. Groups are compared with respect to means on one or more linear composites of the outcome variables. That is, in a MANOVA context, it is the effect of the grouping variable(s) on the linear composite(s) of the outcome variables that is (or should be) of interest to the researcher.

As we indicated in our introduction, all ANOVA-type statistics, require that data conform to distributional assumptions in order to provide valid tests of statistical hypotheses. The validity assumptions for MANOVA include multivariate normality, homogeneity of the P × P covariance matrices, and independence of observations. Empirical findings indicate that when these assumptions are not satisfied rates of Type I and II errors can be seriously distrorted, particularly in nonorthogonal designs (see Christensen & Rencher, 1997; Coombs et al., 1996).

### Research Design Features and Methods of Analysis

What was looked for in the articles reviewed for this content analysis was information related to the conduct of a MANOVA. A summary of some of the information reported for the 79 articles which were examined is given in Table 3.

First, it is sometimes argued by methodologists that, when reasonable, aspects of randomization should be considered in designing a group-comparison study. In only 20 of the 79 studies was randomization considered; 6 involved random selection and 14 involved random assignment. With regard to sample size, one study included an apology for the relatively small sample size used. In another study, it was recognized that "large" Ns were used; therefore, a

relatively low p-value was selected as a cut-off value in determining "significance." Two "conceptually distinct" sets of outcome variables were used in one study; this notion plus the ratios of minimum group size to the number of outcome variables were used by the authors to justify two MANOVAs rather than one MANOVA. [A recommendation that has been proposed is that the smallest group size should range from 6P to 10P (Huberty, 1996).] Statistical power was explicitly addressed in only five articles.

For about 76% (60/79) of the studies, tables of group-by-variable means (and standard deviations) were reported. A matrix of outcome variable intercorrelations was reported in only eight articles.

In an overwhelming 84% (66/79) of the studies, researchers never used the results of the MANOVA(s) to explain effects of the grouping variable(s). Instead, they interpreted the results of multiple univariate analyses. In other words, the substantive conclusions were drawn from the multiple univariate results rather than from the MANOVA. Having found the use of such univariate methods, one may ask: Why were the MANOVAs conducted in the first place? Applied researchers should remember that MANOVA tests linear combinations of the outcome variables (determined by the variable intercorrelations) and, therefore does not yield results that are in any way comparable with a collection of separate univariate tests.

Although it was not indicated in any article, it is surmised that researchers followed the MANOVA-univariate data analysis strategy for protection from excessive Type I errors in the univariate statistical testing. This strategy may not be too surprising because it is suggested by some book authors (e.g., Stevens, 1996, p. 152; Tabachnick & Fidell, 1996, p. 376). There is very limited empirical support for this strategy. A counter position may be stated simply as: Do not conduct a MANOVA unless it is the multivariate effects that are of substantive interest. If the univariate effects are those of interest, then it is suggested that the researcher go directly to the univariate analyses and bypass MANOVA. When doing the multiple univariate analyses, if control over the overall Type I error is of concern (as it often should be), then a Bonferroni (Huberty, 1994, p.17) adjustment or a modified Bonferroni adjustment may be made. (For a more

extensive discussion on the MANOVA versus multiple ANOVAs issue, see Huberty and Morris, 1989.) Focusing on results of multiple univariate analyses preceded by a MANOVA is no more logical than conducting an omnibus ANOVA but focusing on results of group contrast analyses (Olejnik & Huberty, 1993).

If multivariate effects are of interest, then some descriptive discriminant analysis (DDA) techniques would be appropriate (see Huberty, 1994, ch. XV). DDA techniques were used in only four of the 79 studies reviewed. In that one study, four linear discriminant functions were substantively interpreted in discussing group separation. In this same study, techniques of predictive discriminant analysis (PDA) were used "as a descriptive tool to highlight and to further clarify the results (of the DDA)." A second study also mentioned the use of PDA techniques; but by "mixing" PDA and DDA techniques to arrive at classification rules, the analysis lost its meaningfulness.

Assessment of Validity Assumptions

It was disappointing, but perhaps not too surprising, that in only a small percent of the 79 studies were data conditions considered. As indicated, the data conditions of some concern in a MANOVA context pertain to multivariate normality and covariance matrix equality. No studies even mentioned the latter condition. In one study the authors tested for "homogeneity of variances" (which applies only to the univariate context). In six studies, data transformations were used; two studies used the arcsine transformation of proportions and one study used a square root transformation of percents. In one of the repeated measures (RM) MANOVA studies, the condition of sphericity was considered. Very extensive consideration of data conditions was made in one article: normality, covariance matrix homogeneity, sphericity, outliers, covariate regression slopes, and multicollinearity.

Power/Effect Size Analysis

Effect size index values were reported in only eight of the 79 articles. Seven studies used univariate indexes and one study reported multivariate eta-squared values. The actual statistical test criterion (e.g., Wilks) was reported in only a handful of studies; rather, an F value was

reported (usually without any indication of degrees of freedom [df]). All of the four popular test criteria (Bartlett-Pillai, Hotelling-Lawley, Roy, Wilks) may be transformed to F values, so the reporting of an F value does not tell the reader which criterion was used (Huberty, 1994, p. 189). If no criterion value is reported, the reader has some difficulty in arriving at an effect size index value.

Software Packages/Citations

Only 12 of the 79 studies stated the software package used and only 28 of the articles included references to data analysis books and/or articles. This is somewhat surprising considering the data analysis methods used. It may be worth mentioning that even though all 79 articles reviewed were published in 1994 and 1995, some of the data analysis references were not to later editions of books, but rather to editions in the 1980s or before.

Conclusions and Recommendations Concerning Between-Subjects Multivariate Designs

In this section we suggest information that can (should?) be reported in a study that involves a multiple-group, multiple-variable, design in which a MANOVA would be considered.

Pre-Analysis

Outcome variables. Ideally the collection of outcome variables should constitute a variable *system* in the sense that the variables conceptually and substantively "hang together." This initial choice of variables may be based on substantive theory, previous research, expert advice, and professional judgment. The rationale used for including multiple related variables measuring one or more underlying construct(s) should be made clear. Explicit listing (e.g., in a table) of all outcome variables and how each is measured would enhance manuscript readability. Any use of data transformations should be reported. The reporting of the reliability of the measures for each outcome variable would be a real plus.

Outlying observation vectors. As is well known, a few outliers can "foul up" an analysis in surprising ways. An indication that a search for outliers was conducted and steps taken, if any, should be stated. For a discussion of outlier detection in psychology, see Orr, Sackett, and Dubois (1991).

Completeness of data matrix. The manner of handling missing data should be discussed (see for example, Roth, 1994). A second search for outliers may be conducted after the data matrix is completed.

Data conditions. A brief discussion of the extent to which the available data satisfy the conditions of group multivariate normality and equal group covariance matrices should be given. If there is concern about the equality of covariance matrices then various robust alternatives are available (see e.g., Christensen & Rencher, 1997; Coombs et al., 1996; Huberty, 1994, pp. 199, 203). In the two-group problem where $H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ ($\boldsymbol{\mu}_k$ indicates a vector of two or more variable means), researchers can adopt the procedures due to Kim (1992) or Johansen (1980). For the many-group problem where the hypothesis to be tested is $H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_K$, researchers can choose from among the procedures due to Coombs and Algina (in press), James (1954) or Johansen (1980) (see Coombs & Algina, 1996; Coombs et al.). Current findings suggest that for many of the parametric conditions likely to be encountered by behavioral science researchers these procedures should adequately control Type I error; that is, they should provide robust tests of their respective null hypothesis. Assessment of covariance matrix equality and of P-variate normality, including the use of statistical package programs, are discussed by Huberty and Petoskey (in press).

Analysis

Descriptives. There are three basic types of descriptive information for a K-group, P-variable MANOVA situation that should be reported: K means and standard deviations for each outcome variable, and the P × P error correlation matrix. One might also report a K x K matrix of Mahalanobis squared distance values. As a sidenote, another type of information that may be considered *descriptive* consists of the P univariate F values. This descriptive information may indicate to the reader some of the "strong" outcome variables, and, if an F value is less than 1.00, then that variable would be contributing more "noise" than "signal." [Caution: Univariate F tests should not be used to assess relative variable contribution in a multivariate study.]

Statistical tests. For MANOVA main, interaction, or contrast effects, the following test information is suggested: criterion (e.g., Wilks) value, test statistic value (with df values), p-value, and effect size value. Information for contrast effects tests would be the same as for the omnibus effects tests.

Labeling of linear discriminant functions (LDFs). This information would be relevant if an argument is implicitly or explicitly made for approximate equality of group (or cell) covariance matrices. The number of LDFs to consider may be determined in one or more of three ways (statistical tests, proportions of variance, and LDF plots; see Huberty, 1994, pp. 211-216). The retained LDFs may be interpreted/named/labeled by examining the LDF-variable correlations (sometimes called structure rs).

Optional information. Some optional information that may be reported includes LDF plots, outcome variable rank ordering, and outcome variable deletion. These details are reviewed by Huberty (1994, chs. XV, XVI).

## The Analysis of Repeated Measures Designs

Researchers frequently obtain successive measurements from their participants and consequently RM designs often provide the blueprint for experimental manipulations and data collection. RM designs are popular for a number of reasons. First, they are economical in comparison to designs that require an independent group of participants for each treatment combination of independent variables. That is, fewer participants are required in RM designs than completely randomized designs when the effects of certain variables can be measured across the same set of participants. This can be particularly advantageous when participants are expensive to obtain or measure or are scarce in number. A second major advantage of treating a variable as a within-subjects variable as opposed to a between-subjects variable relates to the power to detect treatment effects. By manipulating a variable as a within-subjects variable, that is, by exposing participants to all levels of a variable, variability due to individual differences across the levels of the variable is eliminated from the estimate of error variance thus making it easier to detect treatment effects when they are present. This gain in power can be substantial. Finally, in addition

to economy and sensitivity, RM designs are clearly the design of choice when the phenomenon under investigation is time related, such as when investigating developmental changes, learning and forgetting constructs, or the effects of repeatedly administering a drug or type of therapy.

In this content analysis, three categories were used to define the type of RM research design: simple, single-group factorial, and mixed. In a simple design, a single group of participants is evaluated at each level of one RM factor. In a single-group factorial design, on the other hand, a single group of participants is evaluated at each combination of levels of two or more RM factors. In a mixed design, participants are classified into groups or randomly assigned to groups on the basis of one or more factors and are evaluated at each level of a single RM factor, or at each combination of levels of two or more RM factors. The use of covariates in each of these designs was also noted.

In any of these designs, the conventional ANOVA F test is appropriate for testing RM effects only if the assumption of (multisample) sphericity is met. When sphericity is an untenable assumption, either a df-adjusted univariate approach or a multivariate approach can be adopted. In the former approach, the critical value used in hypothesis testing is based on numerator and denominator df which are modified to reflect the magnitude of the departure from sphericity reflected in the sample data. Two different df-adjusted tests are typically recommended for use by applied researchers, and are often referred to as the Huynh-Feldt and Greenhouse-Geisser (see Maxwell & Delaney, 1990) tests. MANOVA may also be used to test RM effects; this approach does not depend on the sphericity assumption. In designs containing quantitative covariates, the data may be analysed using conventional analysis of covariance (ANCOVA), df-adjusted ANCOVA, or multivariate analysis of covariance (MANCOVA) techniques. For RM designs which are multivariate in nature, and which are analysed as such, multivariate MANOVA or MANCOVA procedures may be used. Multivariate RM data may be analysed from either a multivariate mixed model or doubly multivariate model perspective (Boik, 1988). The former approach assumes that the multivariate (multisample) sphericity assumption is satisfied, while the latter approach does not. Other, less commonly used procedures for testing RM effects include

nonparametric procedures, trend analysis, regression analysis, as well as tests for categorical data such as z tests or chi-square tests of association.

As in between-subjects designs, MCP test statistics that are used in RM designs may be computed in different ways, depending on the assumptions the researcher is willing to make about the data (Keselman & Keselman, 1993). For example, in the simple RM design, one test statistic that may be used incorporates the error term for the omnibus test of the RM effect. As before, we will refer to this as a single-error statistic because the error term is based on the data from all levels of the RM factor. Accordingly, the sphericity assumption must be satisfied for such an approach to provide valid tests of pairwise comparisons (Keselman, 1982). The alternative, a separate-error statistic, uses an error term based on only that data associated with the particular levels of the RM factor that are being compared (Maxwell, 1980). Thus, in the latter approach, which does not depend on the sphericity assumption, each pairwise comparison statistic has a separate-error term. The same concept of single- and separate-error pairwise comparison statistics applies to factorial and mixed RM designs in which multiple within- and/or between-subjects factors exist, but the separate-error statistic may be computed in different ways depending on the assumptions the researcher is willing to make about the data.

Research Design Features and Methods of Analysis

Information pertaining to the classification of the research articles by the type of design is contained in Table 4. Mixed designs were overwhelmingly favored, and were represented in 190 articles (84.1%). Among this number, unbalanced designs (50.5%) were more common than balanced designs (40.5%), although 6 articles reported that both balanced and unbalanced mixed designs were incorporated in a single study (3.2%). Simple designs and single-group factorial designs were rarely used, and were only found in 11.5% and 10% of the articles, respectively.

Total sample size varied considerably across the investigated articles and ranged from six to more than 1000 units of analysis. For mixed designs, 16 articles reported total sample sizes which did not exceed 20 units of analysis, and six reported values greater than 400. However, more than half of the mixed design articles (55.3%) reported total sample sizes of 60 or less units

of analysis. An investigation of group/cell sizes in the articles which contained an unbalanced mixed design revealed that the ratio of the largest to smallest value was not greater than 1.5 in 56.3% of these. Among those articles in which a simple design was used, nine (34.6%) reported a total sample size of 30 units or less, while for the single-group factorial design articles, 14 (63.6%) did so.

Information collected on the types of analyses is also contained in Table 4. As anticipated, inferential techniques were favored in the analysis of all three types of designs, and univariate analyses were more popular than multivariate analyses. In fact, none of the articles relied solely on multivariate techniques for the analysis of RM data; wherever multivariate analyses were performed, they were accompanied by univariate analyses.

Table 5 contains information pertaining to methods of inferential analysis for RM effects. In this table, all of the articles in which the RM factor(s) had only two levels were excluded because in such cases, sphericity is trivially satisfied. If a design employed multiple RM factors, at least one had to have more than two levels in order to be considered in the subsequent analysis. Thus, for mixed, simple, and single-group factorial RM designs, the number of articles that were subjected to analysis were 103, 13, and 12, respectively.

As Table 5 reveals, for mixed designs, the conventional ANOVA F test was overwhelmingly favored (68.9%). A small number of articles (3.9%) reported the use of a mixed design involving covariates for which the authors adopted the conventional ANCOVA F test. In only two mixed design articles was MANOVA used to test RM hypotheses and MANCOVA was used once. In one of the articles in which MANOVA was used, sphericity was evaluated using Mauchly's (1940) test; where a significant result was obtained, a multivariate analysis was adopted instead of the conventional ANOVA approach. In another article where sphericity was tested and a significant result was obtained, both the conservative F test and df-adjusted F test were applied to the data and it was noted whether one or both of the tests were significant. Both multivariate MANOVA (5.8%) and multivariate MANCOVA (1.0%) techniques were used, albiet in a limited manner; the multivariate mixed model perspective was adopted in all of these articles.

In articles where multivariate MANOVA was used, the multivariate analyses were always followed by separate univariate analyses using the conventional ANOVA F test. In the one article where multivariate MANCOVA was used, no univariate tests involving RM effects were conducted; the authors were only interested in univariate tests of between-subject effects.

Six articles reported an incorrect analysis of RM data from mixed designs. In four of these articles, the error df did not correspond to those associated with the reported method of analysis (i.e., ANOVA or MANOVA). In the six articles contained in the not clearly stated category, it was not possible to determine what method of analysis had been used because df were not reported, although it was typically the case that the author(s) stated that an ANOVA approach had been used. Five articles incorporated a mixed design but did not involve an analysis of RM effects; these were classified in the category of no RM analysis.

MPCs of RM means were conducted in almost half of the mixed design articles (see Table 5). It is important to note that given our focus on methods of RM analysis, we did not examine procedures which were used to probe between-subjects effects. The most popular method for RM comparisons was Tukey's procedure, followed by the Newman-Keuls method.

Of those mixed design articles in which pairwise comparisons were performed, marginal means were compared in 25 articles, while simple means were compared in 32 articles. In one of two articles the interaction effect was probed with tetrad contrasts using multiple t-tests. In a two-way design, a tetrad contrast essentially involves testing for the presence of an interaction between rows and columns in a $2 \times 2$ submatrix of the data matrix, and represents a test for a difference in two pairwise differences.

In 43 of the articles in which mean comparisons were performed in mixed designs, it was not clear whether a single- or separate-error test statistic was employed. In seven articles however, a separate-error test statistic was employed.

Table 5 also reports analysis methods for the simple RM designs. Here, use of the conventional ANOVA F test was reported in slightly more than one third of the articles. In six of the 13 simple RM articles, a MCP was used. The Bonferroni and Newman-Keuls procedures were

most popular. In only one article was there an indication that a separate-error test statistic was used in conducting the pairwise comparisons.

Finally, Table 5 reveals that in three-quarters of the single-group factorial studies, the conventional ANOVA approach was used. One of these articles also relied on a df-adjusted ANOVA F test, in this case the Huynh-Feldt correction, when Mauchly's (1940) sphericity test proved to be significant.

Planned contrasts were used in two articles to test specific RM hypotheses in factorial RM designs; in both instances these contrasts followed an omnibus analysis. It is interesting to note that in one of these articles, which involved a $4 \times 3$ single-group factorial design, the test of the interaction effect was followed by a series of $2 \times 3$ planned interaction subanalyses to provide a more specific determination of the source of the interaction.

Pairwise comparisons of means were conducted in one third of those articles in which a single-group factorial RM design was used; information pertaining to the methods adopted is contained in Table 5. It is clear that no one procedure was a clear favorite, as a different method was used in each of the articles. Pairwise comparisons of marginal RM means were reported in three articles, and of simple effect RM means in two. In none of the articles was it possible to discern whether a single- or separate-error test statistic was used.

Assessment of Validity Assumptions

References to problems of distributional assumption violations was evaluated for the entire data base, that is, for all 226 articles which incorporated RM designs. In total, in 35 of these articles (15.5%) the author(s) made reference to some aspect of assumption violations in performing tests of statistical significance. The most commonly mentioned issue was normality (n = 26), although none of these articles made reference to a specific test for normality. Rather, it appears that violations of this assumption were assessed via descriptive techniques. The most common method of dealing with nonnormal data was to transform the scores (n = 10), typically with an arcsin method, although a small number of articles (n = 4) reported that outliers were removed from the distribution of scores prior to analysis. Eleven other articles reported that a

transformation had been applied to the distribution of scores, but gave no rationale for applying the transformation (i.e., these articles did not indicate that the normality assumption appeared untenable). Various other problems with data were mentioned. For example, in one article, Levene's (1960) test was applied to the data due to a concern for variance heterogeneity, but the authors did not evaluate the more complex assumption of (multisample) sphericity.

Power Analysis/Effect Size

Issues of statistical power/effect size were considered in 20 of the 226 articles (8.8%) in the database. In 16 of these articles, effect sizes were calculated, with the most common measure being Cohen's (1988) d statistic. In three articles, the authors mentioned that statistically significant findings may not have been revealed because of potentially low power, but no assessments of power were actually performed.

Statistical Software Packages/Citations

Only ten of the 226 articles in the RM database gave specific information concerning the use of a statistical software package. The SPSS program was favored, and was used in seven of the research reports.

A wide variety of statistical references were found in the 226 RM articles. The two most popular sources were Winer (1971) and Cohen and Cohen (1983), which were each cited five times. The former was typically used as a reference for data transformations while the latter was a reference for various statistical analysis issues in regression and ANOVA. Sources which were used specifically for justification in the choice of a RM analysis technique included McCall and Appelbaum (1973), and Games (1980).

Conclusions and Recommendations Concerning Repeated Measures Designs

Educational researchers make use of RM designs in a variety of contexts, but particularly in the study of developmental changes over time. In these instances, researchers should anticipate the existence of heterogeneous correlations among the repeated measurements, since participant responses that are adjacent in time will typically be more strongly correlated than those which are more distant. The existence of such serial correlation patterns will result in the data violating the

sphericity assumption. It is impossible to evaluate the extent to which sphericity may be violated in behavioral science research, as none of the authors of papers included in this review gave details of this aspect of their data. We recommend, however, that the conventional ANOVA approach for tests of within-subjects effects be avoided because of the problems associated with control of Type I errors under even a minimal degree of nonsphericity (Maxwell & Delaney, 1990, p. 474).

Furthermore, while it is difficult to evaluate the extent to which behavioral science data departs from the more complex assumption of multisample sphericity in mixed designs, we also recommend that the conventional ANOVA approach not adopted in such instances. In particular, tests of within-subjects interaction effects are highly susceptible to increased rates of Type I error when the design is unbalanced and multisample sphericity is not satisfied (see e.g., Keselman, Carriere, & Lix, 1993; Keselman & Keselman, 1993).

Despite the likelihood of the sphericity assumption not holding, this rarely appears to be a concern for educational researchers. Rather, the results of this content analysis suggest that in general, researchers do not give much thought to assumption violations when performing tests of statistical significance, as less than 16% of the papers made reference to this issue. When it was clear that assumptions were considered, normality was more likely to be of concern than sphericity, and transformations were a common way of normalizing the distribution of responses.

It is important to consider distributional assumptions not only when conducting omnibus tests of effects, but also when pairwise comparisons of means or other contrasts are performed on the data. A test statistic that employs an error term which is based on all of the data, in other words a single-error term, is based on the assumption of (multisample) sphericity. Rarely in this content analysis was the choice of a test statistic specified, and thus it was difficult to determine what assumptions the researchers were making about the data.

Furthermore, the general practice among the researchers whose articles were evaluated in this content analysis is to probe interaction effects by conducting tests of simple main effects and pairwise comparisons of simple main effect means. This strategy is inappropriate for evaluating

the nature of an interaction effect (Boik, 1993; Lix & Keselman, 1996; Marascuilo & Levin, 1970) because simple effects are confounded by main effects. Thus, if the hypothesis associated with a simple effect test is rejected, the researcher can not conclude whether the result is due to the presence of an interaction or a consequence of a marginal effect. The correct approach of testing specific contrasts regarding the interaction was rarely seen.

We recommend a number of ways by which behavioral science researchers can improve their analyses of RM data. First, we strongly encourage behavioral science researchers to consider the adoption of analysis methods that are robust to RM assumption violations. Preliminary tests of (multisample) sphericity do not provide a sound basis for a data-analytic decision and should therefore be avoided. Sphericity tests are sensitive to departures from multivariate normality and thus, rejection of the null hypothesis does not necessarily imply that the data are nonspherical (Keselman & Keselman, 1993; Keselman, Rogan, Mendoza, & Breen, 1980; Mendoza, 1980). As well, although transformations may result in a more nearly normal distribution, and may also help to equalize heterogeneous variances (Ekstrom, Quade, & Golden, 1990), these manipulations of the data are not likely to change the correlational structure of the data.

It is apparent that df-adjusted univariate procedures and multivariate procedures are severely underutilized in behavioral science research. We strongly encourage the adoption of these two approaches for analysing RM effects in simple and single-group factorial designs. A number of references are available that can help to demystify these procedures and aid in a decision between them, including Davidson (1972), Keselman and Keselman (1993), Maxwell and Delaney (1990), O'Brien and Kaiser (1985) and Romaniuk, Levin, and Hubert (1977). For these designs we also recommend one of the newest approaches to the analysis of repeated measurements, Boik's (1996) empirical Bayes (EB) approach. The EB approach is a blend of the df-adjusted univariate and the conventional multivariate approaches. The major statistical software packages (e.g., the general linear model and/or multivariate programs from SAS and SPSS) can be used to obtain numerical results for each of these approaches.

Furthermore, for mixed designs, although the adoption of either a df-adjusted univariate or multivariate procedure represents a good first step in terms of obtaining more valid tests of RM hypotheses, a new class of procedures that are not dependent on the multisample sphericity assumption are available and their use is strongly encouraged. Keselman et al. (1993) have shown that an approximate df multivariate solution can provide effective control of the Type I error rate in unbalanced mixed designs, provided that total sample size is sufficiently large. A program written in the SAS/IML language is given by Lix and Keselman (1995) for implementing this solution as well as examples and SAS/IML code demonstrating its use. Other approaches to this problem are discussed by Algina (1994), Algina and Oshima (1994), Keselman and Algina (1997) and Keselman, Algina, Kowalchuk and Wolfinger (1997).

A variety of multiple comparison procedures are available for data that do not satisfy the multisample sphericity assumption. An introductory paper on this topic is Keselman, Keselman, and Shaffer (1991). Current research in this area is discussed by Keselman (1994). As well, Lix and Keselman (1996) provide details of procedures that are appropriate for probing interactions in RM designs. In addition, their program can be used to obtain numerical results. A general discussion of this topic is also provided by Boik (1993).

Current research efforts are being directed towards the development of procedures that control the incidence of Type I errors and provide adequate statistical power when both the normality and sphericity assumptions are violated. Wilcox (1993) considers this problem. As well, it should be noted that new methods for the analysis of RM effects that allow the applied researcher to model and specify the correlational structure of the data are now available in the popular statistical packages (i.e., SASs PROC MIXED; see Keselman, Algina, Kowalchuk, & Algina, in press). However, at present, the limited information on this method suggests that it may be problematic when the wrong covariance structure is selected by the researcher (Keselman et al., in press; 1997).

We recommend that behavioral science researchers give serious though to the value of multivariate analyses, rather than considering individual dependent variables in isolation.

Methods for the analysis of RM data in a multivariate context are discussed by Lix and Keselman (1995) and Keselman and Lix (1997).

## The Analysis of Covariance Designs

ANCOVA has two purposes: First, in experimental studies involving the random assignment of units to conditions, the covariate when related to the response variable, reduces the error variance resulting in increased statistical power and greater precision in the estimation of group effects. Second, in nonexperimental studies where random assignment is not used, the covariate when related to the grouping variable, attempts to control for the confounding effect of the covariate.

A great deal has been written regarding the data assumptions made when using the ANCOVA model including: independence, homoscedasticity, homogeneity of regression slopes, linearity, and conditional normality. Violating the first three assumptions can seriously affect the Type I error rate (Glass, Peckham & Sanders, 1972) particularly when the design is nonorthogonal (e.g., Hamilton, 1977; Levy, 1980).

### Research Design Features and Methods of Analysis

For each journal we examined each article and selected those that reported the use of at least one application of univariate ANCOVA. Regression analyses that referred to some variables as covariates were excluded, as were studies which only reported on a multivariate ANCOVA. Most of the articles reviewed reported the results of several applications of ANCOVA as well as other analytic methods. In total we examined 651 articles and found 45 applications of ANCOVA for a seven percent hit rate. A summary of our findings is provided in Table 6.

All but one of the studies used the individual as the unit of analysis. One study provided training to groups of children and appropriately used the group mean as the unit of analysis. One study analyzed both the individual and subgroups, and two studies were applications of hierarchical linear models (HLM) and considered both individuals and classrooms as the units of analysis.

In the applications of ANCOVA that we reviewed, two thirds of the studies (30) involved nonrandomization of the experimental units. This result supports what many believe, that ANCOVA is underutilized in experimental research (Maxwell, O'Callaghan & Delaney, 1993). In one study the researchers analyzed the data with and without the covariate. When the conclusions were the same, the researchers decided not to report on the details of the ANCOVA. None of the nonexperimental studies recognized the problem of measurement errors nor the fact that all of the confounding variables may not have been controlled. Although explicit causal statements were not made, little effort was made to caution readers not to overinterpret the results.

Many statistics textbooks which present ANCOVA limit their discussions to the one-factor design with a single covariate (e.g., Keppel, 1991; Maxwell & Delaney, 1990). Only the most advanced texts address multiple covariates, factorial and RM designs (Kirk, 1995; Winer, Brown, & Michels, 1991). Even the advanced texts do not discuss in great detail how these analyses might be carried out and interpreted. Among the studies using ANCOVA, over one-third (17) used a factorial design and 11 studies used a mixed model design. Thus almost two-thirds (28) of the studies were multifactor designs. In 19 of the studies multiple covariates were used and in two studies the covariate varied by level of the within-subjects factor.

Twenty-one of the studies had two or more between-group factors (17 factorial and four mixed model designs) and 18 of these studies had unequal and disproportional group sizes. The average group size in the nonorthogonal multi-group analyses equalled 34.5, while for the balanced multi-group studies (3) the average group size was 35.3. For eight of these studies only the total sample size and the number of groups were reported. Twenty-one of the studies involved a single between-group factor (14 oneway and 7 mixed model designs), over seventy-five percent (15) of which had unequal and disproportional group sizes that averaged 37.1 units. The balanced single factor designs had an average of 19.4 observations per group. The inequality of group sizes was not extreme for most cases. Two-thirds of the studies had a ratio of largest to smallest group size of less than two. In the single factor designs the largest ratio of largest to smallest group sizes equalled 8.06, while in the multi-factor designs the largest ratio equalled 5.15.

In mixed model designs only the effects involving the between-subjects factor(s) are adjusted by the covariate when the univariate approach to RM is used for hypothesis testing. No adjustment to the within-subjects is made because the same adjustment is made to all levels of the within-subjects factor(s) unless the covariate varies with the level of the within-subjects factor(s). If an adjustment is desired for the within-subjects factor then the multivariate approach to the analysis of RM is needed (Ceurvost & Stock, 1978). Delaney and Maxwell (1981) point out however that the covariate must be adjusted by the covariate grand mean for the multivariate test to be meaningful. In further clarification of this point, Algina (1982) argued that the mean adjusted covariate is needed only when the covariate is a fixed factor, which is generally not the case, and the meaningfulness of the hypothesis test for between-subjects, within-subjects and their interaction depends on the homogeneity of the within cell slopes.

Only one of the 11 studies using a within-subjects factor cited the Delaney and Maxwell (1981) article and used the mean adjusted covariate. None of the articles stated that they used the multivariate approach to test the within-subjects factor. And only one study commented on the equality of the within group regression slopes.

Twelve of the studies reviewed used a MANCOVA and all of these studies followed a significant multivariate test with a series of the univariate ANCOVA tests. Only the univariate analyses are discussed here.

Twenty-one of the studies had at least three levels of an explanatory variable but only eight studies involved variables having more than three levels. Over half (27) of the studies did not use a contrast procedure because either there were only two levels of the explanatory or grouping variable or there were no differences among the levels of the grouping variable having more than two levels. In two studies contrasts would have been appropriate but were not computed and in one study post hoc tests were computed but not specified.

When a MCP was used, the most common (6) procedure was the multiple t test approach using the pooled within-group variance. Five of these analyses were preceded with an omnibus F test. Two additional studies stated they used Fisher's LSD method but with unequal sample sizes

these analyses were equivalent to multiple t tests. Of the eight studies, five examined all pairwise contrasts, two studies examined a subset of all pairwise contrasts, and one study examined a set of orthogonal contrasts. Textbooks (e.g. Keppel, 1991; Maxwell & Delaney, 1990) generally recommend a Bonferroni adjusted multiple t test procedure or the Bryant and Paulson (1976) procedure if the covariate is considered a random variable and all pairwise contrasts are of interest. None of the studies reported using a Bonferroni adjusted significance level or the Bryant-Paulson procedure, although one study referenced Seaman, Levin, and Serlin (1994) who showed that when df = 2 FWE is controlled. Most of the studies reviewed here involved df $\leq$ 2, in which case a MCP to control the FWE therefore is unnecessary.

Assessment of Validity Assumptions

As we indicated previously, the deleterious effects of assumption violations are exacerbated when group sizes are unequal. The majority of the studies reviewed here involved unequal and disproportional sample sizes. Thirty-four of the studies made no comment at all regarding the sample distributions or any attempt to determine whether it appeared reasonable that the assumptions were met. Only 8 of the studies commented on the homogeneity of regression slope assumption. Six of the studies found no evidence that the assumption was violated, one study found the slopes to differ on only one of the 17 outcomes examined and attributed the result to a Type I error. One study found the slopes to be unequal and proceeded to analyze the data using gain scores. Ignoring the assumption of equal within-group regression slopes is equivalent to assuming that there is no interaction between the covariate and the grouping variable. In factorial designs researchers rarely are willing to assume no interaction between explanatory variables without at least testing that assumption. If the regression slopes are unequal an inappropriate adjustment is made in nonrandomized studies and in experimental studies at a minimum statistical power is lost. But perhaps more importantly the interpretation of the treatment effect is suspect when the interaction is present. Rather than ignoring the interaction hypothesis researchers might consider analyzing the data using methods that do not assume homogeneity of regression as suggested by Rogosa (1980).

Finally, only two studies considered normality and four studies commented on homogeneity of variances. Only one study commented on a search for outliers.

Power/Effect Size Analysis

Surprisingly, only 15 studies reported the adjusted means (it was assumed that reported means were unadjusted unless explicitly stated), 11 studies provided some index of effect size with standardized mean difference being the most popular (7), and none of the studies examined reported results in terms of confidence intervals. Several authors over the past several years (e.g., Carver, 1993; Cohen, 1990; Schmidt, 1992) have recommended that in addition to or instead of tests of statistical significance indices of meaningfulness should also be reported. Some have even recommended abandoning the significance test in favor of effect size indicators and confidence intervals (see, for example, Harlow et al., 1997). In the present sample of studies the behavioral science researchers were either unaware of these recommendations or chose to ignore them.

Software Packages/Citations

Only four of the studies reported the computer package used, two used SPSS and two used HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1989). With a procedure like ANCOVA where programming requires little judgment and programs basically report the same statistics, perhaps identification of the specific package is unnecessary. However, when contrasts are tested not all programs are alike. SPSS, for example, in factorial or RM designs do not compute the Scheffe, Tukey, Bryant and Paulson (1976), or Newman-Keuls MCPs, nor is it possible to compute all possible pairwise contrasts. One wonders then whether these procedures were computed correctly because it requires some computation (Kirk, 1995, p 725) to get the correct standard error. SAS (1990), on the other hand, does compute all pairwise contrasts and complex contrasts can be requested. A Bonferroni adjustment can then be easily made. (SAS also does not compute the Bryant and Paulson statistic.)

Seventeen of the articles referenced statistics texts or methodological articles to support the procedures they used to analyze their data. The most frequently cited statistical reference was the textbook by Kirk (1982); it was cited three times.

Conclusions and Recommendations Concerning Covariance Designs

Our review of 45 articles reporting applications of ANCOVA demonstrates the wide applicability of this analytic technique. The technique has been used across a wide range of disciplines, a variety of age groups, and populations. Although extremely flexible in its application, the 45 studies reviewed here only represent a small percentage of the potential applications. In particular, we found only a small number of applications in experimental studies. Researchers have failed to recognize the potential benefits of reduced error variance to increase statistical power and improve precision. To the extent to which our sample of ANCOVA applications is representative of analytic practice with the technique, it appears to us that most reports of the analyses are inadequate and incomplete.

Although ANCOVA is a versatile analytic tool, it can also be misunderstood, misused, and misinterpreted. Researchers appear to be unaware of, or at least fail to recognize, the assumptions that underlie the statistical models they use. The fact that most of the studies reviewed involved unequal and disproportionate group sizes further raises the concern as to the statistical validity of many research findings. Researchers have generally ignored the interaction effects between the covariate(s) and the grouping variables and have failed to examine residual plots to identify heteroscedasticity and outliers. These preliminary analyses are necessary but they need not require an exhaustive discussion or require extended journal space. When heterogeneity of regression exists researchers should consider adopting the method presented by Rogosa (1980); hopefully, more robust methods will be become available for application to ANCOVA problems. A brief paragraph outlining the procedures used to examine the sample data and a summary of the findings would substantially enhance the credibility of the data analysis. McAuliffe and Dembo (1994) demonstrated how these preliminary analyses may be succinctly reported.

Researchers frequently do not provide adequate descriptive statistics including sample sizes at the smallest group level, pretest means, standard deviations and adjusted postest means. A summary table presenting differences among four groups reported by Steinberg, Lamborn, Darling, Mounts and Dornbusch (1994) is a nice example of how these data might be reported.

Researchers have continued to overrely on hypothesis tests, reporting F ratios and p-values. Effect sizes and confidence intervals have been widely recommended but generally ignored by data analysts. Two exceptions are Simpson, Olejnik, Tam, and Supattathum (1994) who reported standardized mean differences, and Seidman, Allen, Aber, Mitchell and Feinman (1994) who used eta-squared to further explicate their results.

## Summing Up and General Recommendations

Based on our surveys we made specific recommendations to researchers concerning how to improve the statistical analysis practice. In the space remaining we will punctuate our literature reviews of data-analytic practices with: (1) comments directed at several general themes and principles evident in them; and (2) further observations and recommendations related to improving the statistical analyses and reports of behavioral science research data.

### General Themes and Principles

Of the several common themes and priniciples identified in the present set of reviews, three pertain specifically to "assumption validity" concerns. These may be summarized as dos and donts for researchers in the following manner:

1. Be wary. Behavioral science researchers should not automatically conduct a "standard" analysis. Times change, as related both to: (1) the advent of newer, more robust, analytic solutions to assumption-violated data; and to: (2) what is known about the distributional conditions under which a specific statistical test may or may not be appropriate. Conscientious researchers should work hard to be apprised of both those newer developments and those differing conditions (see Wilcox, 1998). Indeed, the text book procedures of the '50s and '60s (e.g., conventional univariate F-tests for the analysis of repeated measurements) have been replaced by more sophisticated analyses (e.g., the EB and mixed-model approaches to the anlaysis of repeated measurements) and reliance on the older methods may lead to misleading or erroneous conclusions.

2. Be more intimate with your data. First, researchers need to: (a) have a clear understanding of the statistical model that underlies their analyses, (b) conduct a careful

preliminary analysis of their data, and (c) provide a detailed report of their analytic results. Unfortunately, many of the articles we reviewed lacked one or more of these conditions. With reference to point (b) researchers need to be more proactive in identifying potential distributional abnormalities in their data by not relying exclusively on summary statistics (e.g., sample means, standard deviations, correlations). Rather, attempts should be made to delve further into one's data [e.g., Exploratory Data Analysis (EDA) techniques such as graphs can be examined, including boxplots, normal probability plots, etc.; see Behrens (1997) for a discussion of EDA]. For data that appear to conform to distributional assumptions, proceed in textbook fashion; but with nonconforming data, give serious consideration to more appropriate alternative analysis procedures of the kind indicated in the present review. In accomplishing this goal researchers should identify the statistical software package (particular programs or procedures) that was used to obtain numerical results (by year or version or release). Numerical results for many of the analyses recommended in our article can be obtained either entirely or in part from the major statistical software packages (e.g., the general linear model and/or multivariate programs from SAS, SPSS, SYSTAT).

3. Don't expect one size to fit all. Each new set of data contains its own distributional idiosyncracies and different analytic tools are required for different types of data. Fortunately, and as was noted previously, both new developments in the statistical literature and the associated computer software are proceeding apace. In fact, it could reasonably be argued (on the basis of Type I error and power characteristic studies) that if a "single size" were to "fit all," that single size should be from the class of lesser-known Welch-based ANOVA alternatives, rather than the standard F test itself (see also Lix & Keselman, 1995). Similarly -- and perhaps surprisingly to many educational researchers -- in the context of RM designs, the standard textbook-recommended univariate F test is a disastrous single size to consider!

Several other data-analysis themes were presented in our article as well. These include the following:

- Researchers should pay greater attention to the "substantive" significance (e.g., Robinson & Levin, 1997; Thompson, 1996) of their research findings -- as reflected by various effect size and strength-of-relationship measures -- rather than simply to the "statistical" significance of their findings. We similarly believe that confidence interval estimates should receive greater use (see e.g., Harlow et al., 1997).

- Researchers should regularly concern themselves with the statistical power characteristics of their studies. Even better, from our perspective, researchers should plan their studies (in terms of appropriate sample sizes) so as to have sufficient power to detect effects that are deemed to be of substantive importance (e.g., Cohen, 1988; Levin, 1997).

- Researchers should similarly think about their specific research questions prior to conducting their studies so that they can select the most appropriate and powerful analytic techniques by which to analyze their data (e.g., Levin, 1998; Marascuilo & Levin, 1970). Omnibus hypothesis tests should not be routinely conducted when individual contrasts form the basis of the researcher's major questions of interest. Multivariate tests should generally be reserved for questions about multivariate structure. Thus, researchers need to translate their research questions into specific and detailed statistical hypotheses.

- Researchers should avoid making "logical inconsistency" errors (or what have been called Type IV errors (see Marascuilo & Levin, 1970) in their analyses. Incorrect interpretations of rejected interaction hypotheses constitute but one salient example of this type of problem that was encountered in the surveyed studies.

We, of course, know that there are a number of practical issues that affect research practice, and in particular, the manner in which data are analyzed. These include: (1) the (limited) training that occurs in graduate-level statistical methods courses; (2) the views of journal editors regarding the types of analyses that they believe are appropriate; (3) the restricted ability of researchers to hire statistical consultants on their projects; (4) the inaccessibility and/or complexity of statistical software; and (5) the cultural milieu within the present-day educational research community. We consider each of these obstancles in turn.

First, with regard to graduate-level training, we have noted that quantitative methods courses have diminished in the set of students' required/recommended courses in many of our graduate programs. When such courses are included in the curriculum, they are frequently taught by colleagues whose speciality is not quantitaive methods. We consider such circumstances to be unfortunate and disadvantageous to students whose careers will involve either conducting empirical research or consuming empirical research findings.

Second, editors of professional journals obviously have their own biases regarding the "proper" analyses that should accompany research reports, as well as the ones they would prefer to see in their own journals. We can only hope that editors, in addition to the researchers themselves, will take notice of the points raised in our review.

Third, in this era of dwindling financial resources for educational research, possibilities for allocating funds for statistical consultation have similarly dwindled. Nonetheless, with whatever funds are available, educational researchers should consider adopting the medical model where having a statistical consultant on board is common research practice.

Fourth, in reference to the inaccessibility/complexity of noncommercially produced software of the kind recommended in this article (e.g., Lix and Keselman's, 1995 SAS/IML program for obtaining robust analyses, particularly in repeated measures designs], we note that they are becoming more accessible, almost on a daily basis, through the internet and its downloading facility. Some might argue that using such programs is beyond the capability of researchers who are not quantitative experts. We, on the other hand, do not subscribe to that position but instead maintain a "let's see" attitude. Frankly, we do not think our profession is well served if the newest developments in an area are hidden simply because of the fear that colleagues might find those developments challenging.

Finally, what about the cultural milieu in educational research today? The appropriate-statistical-analysis message delivered in this article might seem like very "small potatoes" indeed in a field that is currently struggling with more overreaching philosophical issues, such as the role and importance of quantitative methods <u>at all</u> in educational research. Why then are we so

concerned about what seem to be much less important and esoteric matters as distributional assumptions and the validity of statistical tests? Obviously, one isolated article, with its restricted focus, cannot resolve the quantitative-qualitative debate. In fact, the present article was not intended even to discuss it. Our purpose here was to argue that if and when inferential statistical methods are the analytic tools of choice, then at least those tools should be used wisely and properly, in a "statistically valid" (Cook & Campbell, 1979) way. Improper use is likely to lead to danger, in the form of researcher conclusions that are unwaranted on the basis of the evidence presented and analyzed. Consequently, our plea to educational researchers is twofold: (a) be more concerned about mismatches between your evidence and the conclusions you reach; and (b) seek out and embrace statistical methods that are known to reduce that mismatch.

In conclusion, this review should serve as an wake-up call to substantive and quantitative researchers alike. Substantive researchers need to wake up both to the (inappropriate) statistical techniques that are currently being used in practice and to the (more appropriate) ones that should be being used. Quantitative researchers need to wake up to the needs of substantive researchers. If the best statistical developments and recommendations are to be incorporated into practice, it is critical that quantitative researchers broaden their dissemination base and publish their findings in applied journals in a fashion that is readily understandable to the applied researcher.

## Footnotes

This research was supported in part by a grant from the Social Sciences and Humanities Council of Canada. Authorship is listed alphabetically within each tier.

1. The content analyses that follow were originally presented as a symposium at the 1996 annual meeting of The Psychometric Society in Banff, Canada. The title and authors of those papers were: (1) The analysis of between-subjects univariate designs (Lix, Cribbie, & Keselman, 1996), (2) The analysis of between-subjects multivariate designs (Huberty & Lowman, 1996), (3) The analysis of repeated measures designs (Kowalchuk, Lix, & Keselman, 1996), and (4) The analysis of covariance designs (Olejnik & Donahue, 1996). The symposium concluded with a discussion by Joanne C. Keselman and Joel R. Levin.

References

Algina, J. (1982). Remarks on the analysis of covariance in repeated measures designs. Multivariate Behavioral Research, 17, 117-130.

Algina, J. (1994). Some alternative approximate tests for a split plot design. Multivariate Behavioral Research, 29, 365-384.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. British Journal of Mathematical and Statistical Psychology, 47, 151-165.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, D.C.: American Psychological Association.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2, 131-160.

Boik, R. J. (1988). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. Communications in Statistics, Theory and Methods, 20, 1233-1255.

Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. Journal of Educational Statistics, 18, 1-40.

Boik, R. J. (1997). Analysis of repeated measures under second-stage sphericity: An empirical Bayes approach. Journal of Educational and Behavioral Statistics, 22, 155-192.

Bryant, J. L., & Paulson, A. S. (1976). An extension of Tukey's method of multiple comparisons to experimental designs with random concomitant variables. Biometrika, 63, 631-638.

Bryk, A., Raudenbush, S., Seltzer, M. & Congdon, R., Jr. (1989). An introduction to HLM: Computer program and users' guide. Chicago: University of Chicago Press.

Carlson, J. E., & Timm, N. H. (1974). Analyses of nonorthogonal fixed effects designs. Psychological Bulletin, 81, 563-570.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Ceurvorst, R, W., & Stock, W. A. (1978). Comments on the analysis of covariance with repeated measures designs. Multivariate Behavioral Research, 13, 509-513.

Christensen, W. F., & Rencher, A. C. (1997). A comparison of Type I error rates and power levels for seven solutions to the multivariate Behrens-Fisher problem. Communications in Statistics-Simulation, 26, 1251-1273.

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

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

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

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally.

Coombs, W. T., Algina, J. (in press). New test statistics for MANOVA/descriptive discriminant analysis. Educational and Psychological Measurement.

Coombs, W. T., Algina, J. (1996). On sample size requirements for Johansen's test. Journal of Educational and Behavioral Statistics, 21, 169-178.

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. Review of Educational Research, 66, 137-179.

Coursol, A., & Wagner, E. E. (1986). Affect of positive findings on submission and acceptance rates: A note on meta-analysis bias. Professional Psychology, 17, 136-137.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. Psychological Bulletin, 77, 446-452.

Delaney, H. D., & Maxwell, S. E. (1981). On using analysis of covariance in repeated measures design. Multivariate Behavioral Research, 16, 105-124.

Edgington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.

Ekstrom, D., Quade, D., & Golden, R. N. (1990). Statistical analysis of repeated measures in psychiatric research. Archives of General Psychiatry, 47, 770-772.

Elmore, P. B., & Woehlke, P. L. (1988). Statistical methods employed in *American Educational Research Journal*, *Educational Researcher*, and *Review of Educational Research* from 1978 to 1987. Educational Researcher, 17(9), 19-20.

Elmore, P. B., & Woehlke, P. L. (1998, April). Twenty years of research methods employed in *American Educational Research Journal*, *Educational Researcher*, and *Review of Educational Research*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Games, P. A. (1980). Alternative analyses of repeated-measure designs by ANOVA and MANOVA. In A. Von Eye (Ed.), Statistical methods in longitudinal research: Vol. 1, Principles and structuring change (pp. 81-121). San Diego, CA: Academic Press.

Glass, G. V, Peckham P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research, 42, 237-288.

Goodwin, L. D., & Goodwin, W. L. (1985a). An analysis of statistical techniques used in the *Journal of Educational Psychology*, 1979-1983. Educational Psychologist, 20, 13-21.

Goodwin, L. D., & Goodwin, W. L. (1985b). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.

Hamilton, B. L. (1977). An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance. Educational and Psychological Measurement, 37, 701-712.

Harlow, L. L., Muliak, S. A., & Steiger, J. H. (1997). (Eds.). What if there were no significance tests? Hillsdale, NJ: Erlbaum.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

Hsiung, T., & Olejnik, S. (1994a). Contrast analyses for additive nonorthogonal two-factor design in unequal variance cases. British Journal of Mathematical and Statistical Psychology, 47, 337-354.

Huberty, C. J (1994). Applied discriminant analysis. New York: Wiley.

Huberty, C. J (1996, August). Some issues and problems in discriminant analysis. Paper presented at the Joint Statistical Meetings, Chicago.

Huberty, C. J, & Lowman, L. L. (1996, June). The analysis of multivariate designs. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.

Huberty, C. J, & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.

Huberty, C. J, & Petoskey, M. D. (in press). Multivariate anlaysis of variance and covariance. In H. E. A. Tinsley and Brown, S. (eds), Handbook of multivariate statistics and mathematical modeling. San Diego, CA: Academic Press.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biometrika, 41, 19-43.

Keppel, G. (1991). Design and Analysis A Researcher's Handbook (3rd ed.). Englewood Cliffs: Prentice-Hall.

Keselman, H. J. (1982). Multiple comparisons for repeated measures means. Multivariate Behavioral Research, 17, 87-92.

Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. Journal of Educational Statistics, 19, 127-162.

Keselman, H. J., & Algina, J. (1997). The analysis of higher-order repeated measures designs. In B. Thompson (Ed.), Advances in social science methodology. Greenwich, CT: JAI Press.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (in press). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Communications in Statistics – Simulation and Computation.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1997). The analysis of repeated measurements with mixed-model Satterthwaite F tests. Unpublished manuscript.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. Journal of Educational Statistics, 18, 305-319.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. Psychometrika, 60, 395-418.

Keselman, H. J., & Keselman, J. C. (1993). Analysis of repeated measurements. In L. K. Edwards (Ed.) Applied analysis of variance in behavioral science (pp. 105-145). New York: Marcel Dekker.

Keselman, H. J., Keselman, J. C., & Games, P. A. (1991). Maximum familywise Type I error rate: The least significant difference, Newman-Keuls, and other multiple comparison procedures. Psychological Bulletin, 110, 155-161.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. Psychological Bulletin, 110, 162-170.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. Psychometrika, 63, 45-163.

Keselman, H. J., & Lix, L. M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. British Journal of Mathematical and Statistical Psychology, 50, 319-338.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1997). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.

Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures F tests. Psychological Bulletin, 87, 479-481.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed). Belmont, CA: Brooks/Cole.

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed). Belmont, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Kowalchuk, R., K., Lix, L. M., & Keselman, H. J. (1996, June). The analysis of repeated measures designs. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics. Stanford: Stanford University Press.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. Psychology and Aging, 12, 84-106.

Levin, J. R. (1998). To test or not to test $H_0$? Educational and Psychological Measurement, 58, 313-333.

Levin, J. R., Serlin, R.C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. Psychological Bulletin, 115, 153-159.

Levy, K. J. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. Educational and Psychological Measurement, 40, 835-840.

Lix, L. M., Cribbie, R., & Keselman, H. J. (1996, June). The analysis of between-subjects univariate designs. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. Psychological Bulletin, 117, 547-560.

Lix, L. M., & Keselman, H. J. (1996). Interaction contrasts in repeated measures designs. British Journal of Mathematical and Statistical Psychology, 49, 147-162.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 58, 409-429.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Review of Educational Research, 66, 579-620.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. American Educational Research Journal, 7, 397-421.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics, 5, 269-287.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

Maxwell, S. E., O'Callaghan, M. F., & Delaney, H. D. (1993). Analysis of Covariance. In L. K. Edwards (Ed.) Applied analysis of variance in behavioral science. New York:Dekker.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. Annals of Mathematical Statistics, 29, 204-209.

McAuliffe, T. J., & Dembo, M. H. (1994). Status rules of behavior in scenarios of peer learning. Journal of Educational Psychology, 86(2), 163-172.

McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. Child Development, 44, 401-415.

Mendoza, J. L. (1980). A significance test for multisample sphericity. Psychometrika, 45, 495-498.

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. Psychological Bulletin, 101, 464-470.

Norusis, M. J. (1993). SPSS for windows advanced statistics release 6. Chicago, Ill: SPSS Inc.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. Psychological Bulletin, 97, 316-333.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for t-tests through multivariate hypotheses. In L. K. Edwards (Ed.), Applied analysis of variance in behavioral science (pp. 297-344). New York: Marcel Dekker.

Olejnik, S., & Donahue, B. (1996, June). The analysis of covariance designs. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.

Olejnik, S., & Hess, B. (1997). Top ten reasons why most omnibus ANOVA F-tests should be abandoned. Journal of Vocational Education Research, 22, 219-232.

Olejnik, S., & Huberty, C. J (1993, April). Preliminary statistical tests. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Olejnik, S., & Lee, J. L. (1990). Multiple comparison procedures when population variances differ . University of Georgia. (ERIC Document Reproduction Service No. ED 319 754).

Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in "I/Opsychology": A survey of researcher beliefs and an empirical illustration. Personnel Psychology, 44, 473-486.

Ridgeway, V. G., Dunston, P. J., & Qian, G. (1993). A methodological analysis of teaching and learning strategy research at the secondary school level. Reading Research Quarterly, 28, 335-349.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26, 21-26.

Rogosa, D. (1980). Comparing non-parallel regression lines. Psychological Bulletin, 88, 307-321.

Romaniuk, J. G., Levin, J. R., & Hubert, L. J. (1977). Hypothesis-testing procedures in repeated measures designs: On the road map not taken. Child Development, 48, 1757-1760.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-560.

SAS (1990). SAS/STAT user's guide. Cary, NC: Author.

Schmidt, F. L. (1992). What do data really mean? American Psychologist, 47, 11873-1181.

Seaman, M. A, Levin, J. R., & Serlin, R. C. (1994). A controlled, powerful multiple-comparison strategy for several situations. Psychological Bulletin, 115, 153-159.

Seidman, E., Allen, L., Aber, J. L., Mitchell, C., & Feinman, J. (1994). The impact of school transitions in early adolescence on the self-esteem and perceived social context of poor urban youth. Child Development, 65, 507-522.

Simpson, M. L., Olejnik, S., Tam, A. Y., & Supattathum, S. (1994). Elaborative verbal rehearsals and college students' cognitive performance. Journal of Educational Psychology, 86(2), 267-278.

Steinberg, L., Lamborn, S. D., Darling, N., Mounts, N. S., & Dornbusch, S. M. (1994). Over-time changes in adjustment and competence among adolescents from authoritative, authoritarian, indulgent, and neglectful families. Child Development, 65, 754-770.

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

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics. (3rd ed). New York: Harper Collins.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

West, C. K., Carmody, C., & Stallings, W. M. (1983). The quality of research articles in the *Journal of Educational Research*, 1970 and 1980. Journal of Educational Research, 77, 70-76.

Wilcox, R. R. (1987). New designs in analysis of variance. Annual Review of Psychology, 38, 29-60.

Wilcox, R. R. (1993). Analysing repeated measures or randomized block designs using trimmed means. British Journal of Mathematical and Statistical Psychology, 46, 63-76.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65, 51-77.

Wilcox, R. R. (1996). Statistics for the social sciences. New York: Academic Press.

Wilcox, R. R. (1998). How may discoveries have been lost by ignoring modern statistical methods? American Psychologist, 53, 300-314.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and $F^*$ statistics. Communications in Statistics – Simulation and Computation, 15, 933-943.

Wilkinson, L. (1988). SYSTAT: The System for statistics. Evanston, IL: SYSTAT Inc.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design. New York: McGraw Hill.

Table 1. Journal Source and Frequency for the Content Analyses

| Journal | BSUD | BSMD | RMD | CD |
|---|---|---|---|---|
| *American Educational Research Journal* | 4 | 4 | 5 | |
| *Child Development* | 16 | 34 | 56 | 10 |
| *Cognition and Instruction* | 3 | | 5 | 1 |
| *Contemporary Educational Psychology* | 5 | | 19 | 3 |
| *Developmental Psychology* | 7 | 12 | 52 | 5 |
| *Educational Technology, Research and Development* | 1 | | 1 | |
| *Journal of Applied Psychology* | | 10 | | |
| *Journal for Research in Mathematics Education* | | | 3 | |
| *Journal of Counseling Psychology* | 3 | 10 | 10 | 2 |
| *Journal of Educational Computing Research* | 10 | | 17 | 6 |
| *Journal of Educational Psychology* | 6 | 9 | 20 | 7 |
| *Journal of Experimental Child Psychology* | 5 | | 33 | 1 |
| *Journal of Experimental Education* | | | | 3 |
| *Journal of Personality and Social Psychology* | | | | 6 |
| *Journal of Reading Behavior* | | | 3 | |
| *Reading Research Quartely* | | | | 1 |
| *Sociology of Education* | 1 | | 2 | |
| ***TOTAL*** | **61** | **79** | **226** | **45** |

Note: BSUD=Between-Subjects Univariate Design; BSMD=Between-Subjects Multivariate Design; RMD=Repeated Measures Design; CD=Covariance Design.

Table 2. Between-Subjects Univariate Design and Methods of Analysis

| Variable | n | % |
|---|---|---|
| Research Design[a] | | |
| One-way | 36 | 59.0 |
| Balanced Only | 13 | 36.1 |
| Unbalanced Only | 21 | 58.3 |
| Both Balanced & Unbalanced | 2 | 5.6 |
| Factorial | 29 | 47.5 |
| Orthogonal Only | 7 | 24.1 |
| Nonorthogonal Only | 21 | 72.4 |
| Not Stated | 1 | 3.4 |
| Inferential Analysis Techniques (n = 60)[b] | | |
| ANOVA F | 56 | 93.3 |
| Nonparametric | 4 | 6.7 |
| Planned Contrasts | 2 | 3.3 |
| Trend Analysis | 1 | 1.7 |
| Incorrect Analysis | 1 | 1.7 |
| No Omnibus Analysis | 2 | 3.3 |
| Pairwise Multiple Comparisons (n = 29)[c] | | |
| Tukey | 8 | 27.6 |
| Newman-Keuls | 6 | 20.7 |
| Scheffe | 4 | 13.8 |
| Fisher LSD | 3 | 10.3 |
| Multiple t-tests | 3 | 10.3 |
| Duncan | 2 | 6.9 |
| Bonferroni | 1 | 3.4 |
| Hayter | 1 | 3.4 |
| Nonparametric | 1 | 3.4 |
| Spjotvoll & Stoline | 1 | 3.4 |
| Not Stated | 1 | 3.4 |

[a]Totals may not sum to 61 and percentages may not sum to 100 because some articles were included in more than one category.
[b]Totals may not sum to 60 and percentages may not sum to 100 because some articles were included in more than one category.
[c]Totals may not sum to 29 and percentages may not sum to 100 because some articles were included in more than one category.

Table 3. Frequencies of Information Reported in Journal Articles for Between-Subjects Multivariate Designs

| Information Reported | AERJ (4) | CD (34) | DP (12) | JAP (10) | JCP (10) | JEP (9) | T (79) |
|---|---|---|---|---|---|---|---|
| Year | | | | | | | |
| 1994 | 0 | 12 | 4 | 7 | 0 | 2 | 25 |
| 1995 | 4 | 22 | 8 | 3 | 10 | 7 | 54 |
| Covariance Matrix Equality | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Other Data Conditions | 0 | 8 | 0 | 0 | 0 | 1 | 9 |
| Data Transformations | 0 | 6 | 0 | 0 | 0 | 0 | 6 |
| Power/Sample Size Considerations | 0 | 0 | 3 | 0 | 2 | 0 | 5 |
| Table of Means | 4 | 25 | 8 | 8 | 9 | 6 | 60 |
| Correlation Matrix | 0 | 5 | 2 | 2 | 5 | 1 | 15 |
| Multiple Univariate Analyses | 4 | 28 | 9 | 8 | 8 | 9 | 66 |
| Multiple-Factor Design | 2 | 24 | 7 | 7 | 5 | 7 | 52 |
| MANOVA + PDA | 0 | 1 | 0 | 3 | 0 | 0 | 4 |
| MANOVA + DDA | 0 | 1 | 0 | 3 | 0 | 0 | 4 |
| Effect Size | 0 | 3 | 2 | 1 | 1 | 1 | 8 |
| Computer Package | 1 | 3 | 1 | 3 | 2 | 2 | 12 |
| Data Analysis Reference(s) | 2 | 6 | 4 | 6 | 4 | 6 | 28 |
| Randomization | | | | | | | |
| Selection | 0 | 4 | 2 | 0 | 0 | 0 | 6 |
| Assignment | 0 | 3 | 3 | 4 | 2 | 2 | 14 |

Note: Parentheses following journal abbreviation enclose number of articles reviewed.AERJ - *American Educational Research Journal*, CD- *ChildDevelopment*,DP- *Developmental Psychology*, JAP- *Journal of Applied Psychology*, JCP- *Journal of Counseling Psychology*, JEP- *Journal of Educational Psychology,* T-All.

Table 4. Research Design and Analysis for Repeated Measures Designs

| **Variable** | n | % |
|---|---|---|
| Research Design[a] | | |
| Mixed | 190 | 84.1 |
| Unbalanced Only | 96 | 50.5 |
| Balanced Only | 77 | 40.5 |
| Balanced & Unbalanced | 6 | 3.2 |
| Not Clearly Stated | 10 | 5.3 |
| N/A (continuous independent variables) | 1 | .5 |
| Simple | 26 | 11.5 |
| Single-Group Factorial | 22 | 9.7 |
| Type of Analysis | | |
| Mixed | | |
| Univariate | 156 | 86.7 |
| Univariate and Multivariate | 24 | 13.3 |
| Descriptive Only | 10 | 5.3 |
| Simple | | |
| Univariate | 21 | 100.0 |
| Descriptive Only | 5 | 19.2 |
| Single-Group Factorial | | |
| Univariate | 17 | 94.4 |
| Univariate and Multivariate | 1 | 5.6 |
| Descriptive Only | 4 | 18.2 |

[a]Some articles used designs in more than one category, therefore frequencies do not sum to 226 and percentages do not sum to 100.

Table 5. Methods of Inferential Analysis for Repeated Measures Designs

| Variable | Mixed n | Mixed % | Simple n | Simple % | Factorial n | Factorial % |
|---|---|---|---|---|---|---|
| Analysis Methods[a] | n = 103 | | n = 13 | | n = 12 | |
| Conventional ANOVA F | 71 | 68.9 | 5 | 38.5 | 9 | 75.0 |
| Conventional ANCOVA F | 4 | 3.9 | -- | -- | -- | -- |
| MANOVA | 2 | 1.9 | -- | -- | -- | -- |
| MANCOVA | 1 | 1.0 | -- | -- | -- | -- |
| DF-Adjusted ANOVA F | 2 | 1.9 | -- | -- | 1 | 8.3 |
| Conservative ANOVA F | 1 | 1.0 | -- | -- | -- | -- |
| Multivariate MANOVA | 6 | 5.8 | -- | -- | -- | -- |
| Multivariate MANCOVA | 1 | 1.0 | -- | -- | -- | -- |
| Planned Contrasts | 4 | 3.9 | 1 | 7.7 | 2 | 16.7 |
| Trend Analysis | 5 | 4.8 | -- | -- | -- | -- |
| Regression | 3 | 2.9 | -- | -- | -- | -- |
| Nonparametric | 1 | 1.0 | -- | -- | -- | -- |
| Frequency Analysis | 6 | 5.8 | -- | -- | 1 | 8.3 |
| Correlation | 14 | 13.6 | 4 | 30.8 | 3 | 25.0 |
| Incorrect Analysis | 6 | 5.8 | 3 | 23.1 | -- | -- |
| Not Clearly Stated | 6 | 5.8 | 1 | 7.7 | -- | -- |
| No RM Analysis | 5 | 4.8 | -- | -- | -- | -- |
| Other Analysis | 1 | 1.0 | 3 | 23.1 | -- | -- |
| Pairwise Multiple Comparisons[a] | n = 50 | | n = 6 | | n = 4 | |
| Tukey | 15 | 30.0 | -- | -- | 1 | 25.0 |
| Newman-Keuls | 9 | 18.0 | 2 | 33.3 | -- | -- |
| Multiple t-tests | 8 | 16.0 | 1 | 16.7 | -- | -- |
| Bonferroni | 6 | 12.0 | 2 | 33.3 | 1 | 25.0 |
| Scheffe | 5 | 10.0 | -- | -- | -- | -- |
| Fisher LSD | 2 | 4.0 | 1 | 16.7 | -- | -- |
| Duncan | 1 | 2.0 | -- | -- | -- | -- |
| Not Stated | 3 | 6.0 | 1 | 16.7 | 1 | 25.0 |
| Other | 2 | 4.0 | -- | -- | 1 | 25.0 |

[a]Frequencies may not sum to n and percentages may not sum to 100 because some articles reported more than one analysis technique.

Table 6. Summary of ANCOVA applications

| Design Characteristics | | | | | |
|---|---|---|---|---|---|
| Unit of Analysis | n | Independent Variables | n | Sample Size[1] | n |
| Individual | 41 | One Factor | 14 | One Grouping Factor | |
| Group | 1 | Factorial | 17 | Balanced | 5 |
| Both | 3 | 2 X 2 | 4 | Unbalanced | 15 |
| | | 2 X 3 | 6 | Factorial | |
| | | 3 X 5 | 1 | Balanced | 3 |
| | | 4 X 4 | 1 | Unbalanced | 18 |
| | | Higher | 5 | | |
| | | RM | 1 | Covariates | |
| | | Mixed Model | 11 | Single | 26 |
| | | One Between | 7 | Multiple | 19 |
| | | Two Between | 4 | | |
| | | 2 X 2 | 3 | | |
| | | 2 X 3 | 1 | | |
| | | Hierarchial | 2 | | |
| Analyses/Software/Assumptions | | | | | |
| MCPs | n | Computer Programs | n | Validity Assumptions | n |
| Not Applicable | 27 | HLM | 2 | No Comment | 34 |
| Multiple t tets | 6 | SPSS | 2 | Variances | 4 |
| Newman-Keuls | 4 | Not Specified | 41 | Normality | 2 |
| Scheffe | 2 | | | Slopes | 8 |
| Fisher | 2 | | | | |
| Tukey | 1 | | | | |
| No test | 2 | | | | |
| ? | 1 | | | | |

[1]Note: Does not include two HLM applications, one RM analysis without grouping, and one analysis using groups.