ED 318 753 TM 014 770

AUTHOR Thompson, Bruce
TITLE Planned versus Unplanned and Orthogonal versus Nonorthogonal Contrasts: The Neo-Classical Perspective.
PUB DATE 30 Mar 90
NOTE 49p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; Comparative Analysis; Error of Measurement; Hypothesis Testing; Statistical Significance
IDENTIFIERS Nonorthogonal Variables; *Orthogonal Comparison; *Planned Comparisons

ABSTRACT

The use of multiple comparisons in analysis of variance (ANOVA) is discussed. It is argued that experimentwise Type I error rate inflation can be serious and that its influences are often unnoticed in ANOVA applications. Both classical balanced omnibus and orthogonal planned contrast tests inflate experimentwise error to an identifiable maximum. Significance test results are overinterpreted in contemporary analytic practice, and researchers must consider effect sizes and replicability or invariance analyses when formulating interpretations. To guide analytic practice, it is suggested that omnibus hypotheses that are not of interest or which cannot be interpreted should not be tested, since such tests can distort hypothesis tests that are of interest. Orthogonal contrasts should be preferred over non-orthogonal contrasts. The use of planned contrasts is suggested in place of omnibus or unplanned hypothesis tests. Use of planned comparisons tends to result in more thoughtful research with greater power against Type II error. Small data sets and examples support the discussion, and nine tables illustrate these examples. A 73-item list of references is included. (SLD)

aerapowr.wp 3/30/90

PLANNED VERSUS UNPLANNED AND ORTHOGONAL VERSUS NONORTHOGONAL CONTRASTS: THE NEO-CLASSICAL PERSPECTIVE

Bruce Thompson
University of New Orleans 70148

Paper presented at the annual meeting of the American Educational Research Association, Boston, April 19, 1990.

ABSTRACT

The literature regarding the use of multiple comparisons in analysis of variance is reviewed. Three analytic premises provide a framework for the discussion. It is argued that experimentwise Type I error rate inflation can be serious, and that its influences are often unrecognized in many ANOVA applications. It is noted that both classical balanced omnibus and orthogonal planned contrast tests inflate experimentwise error to an identifiable maximum. Finally, it is suggested that significance test results are overinterpreted in contemporary analytic practice, and that researchers must also consider effect sizes and replicability or invariance analyses when formulating interpretations.

Three canons to guide analytic practice are suggested. First, it is suggested that omnibus hypotheses that are not of interest or which cannot be interpreted should not be tested, since such tests can distort the hypothesis tests that are of interest, as illustrated with examples. Second, it is suggested that orthogonal contrasts should be preferred over nonorthogonal contrasts, and that when nonorthogonal contrasts must be used it may be necessary to use a corrected testwise alpha level. Finally, it is suggested that planned contrasts should be used in place of either omnibus or unplanned hypothesis tests. Two reasons why planned comparisons are generally superior are presented. Use of planned comparisons tends to result in more thoughtful research with greater power against Type II error. Throughout the paper small data sets and examples are employed to make the discussion concrete.

Empirical studies of research practice (Edgington, 1974; Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980) indicate that the classical analysis of variance (ANOVA) methods presented by Fisher (1925) several generations ago remain popular with social scientists, notwithstanding withering criticisms of some of these applications (Cohen, 1968; Thompson, 1986). Most users of ANOVA-type methods (ANOVA, ANCOVA, MANOVA, MANCOVA--hereafter labelled OVA methods) are aware that "A researcher cannot stop his analysis after getting a significant _F_; he must locate the cause of the significant _F_" for an omnibus test (Huck, Cormier & Bounds, 1974). An omnibus test evaluates differences across all groups in the way or effect as a set, and has degrees of freedom equal to those available for the effect (e.g., in a 4x3 design the omnibus test for the four-level "A" way has 4-1 or 3 degrees of freedom). Gravetter and Wallnau (1985, p. 423) concur that "Reject Ho indicates that at least one difference exists among the treatments. With _k_ [means] = 3 or more, the problem is to find where the differences are." Moore (1983, p. 299) suggests that:

> If we have statistical significance when we have only two groups, and thus only two means, we can visually inspect the data to determine which group performed better than the other. But when we have three or more groups, we need to investigate specific mean comparisons.

Many researchers employ unplanned (also called a posteriori or post hoc) multiple comparison tests (e.g., Scheffe, Tukey, or Duncan) to isolate means that are significantly different within OVA ways (also called factors) having more than two levels. As Glass and Hopkins (1984, p. 368) note,

> MC procedures are a relatively recent addition to the statistical arsenal; most MC techniques were developed during the 1950's, although their use in behavioral research was rare prior to the 1960's.

Textbook authors tend to discuss unplanned comparison or contrast procedures in a somewhat pejorative terms. For example, Kirk (1984, p. 360) speaks of the use of unplanned comparisons as "ferreting out significant differences among means, or, as it is often called, data snooping." The following quotations are additional representatives of this genre of views:

> Techniques that have been developed for data snooping following an over-all [significant omnibus] *F* test... are referred to as *a posteriori* or *post hoc* tests. (Kirk, 1968, p. 73)

> The post hoc method is suited for trying out hunches gained during the data analysis. (Hays, 1981, p. 439)

> Post hoc comparisons, on the other hand, enable the researcher to engage in so-called data snooping by performing any or all of the conceivable comparisons

> between means. (Pedhazur, 1982, p. 305)

> Prior to running the experiment, the investigator in our example had no well-developed rationale for focusing on a particular comparison between means. His was a "fishing expedition"... Such comparisons are known as post hoc comparisons, because interest in them is developed "after the fact"--it is stimulated by the results obtained, not by any prior rationale. (Minium & Clarke, 1982, p. 321)

> Post hoc comparisons often take the form of an intensive "milking" of a set of results--e.g., the comparison of all possible pairs of treatment means. (Keppel, 1982, p. 150)

> Post hoc comparisons are made in accordance with the serendipity principle--that is, after conducting your experiment you may find something interesting that you were not initially looking for. (McGuigan, 1983, p. 151)

Planned (also called a priori or focused) comparisons provide an alternative to the OVA user who is interested in isolating differences among means. As Keppel (1982, p. 164) notes in his excellent treatment, decisions about which unplanned or planned comparisons to employ in OVA research are complex and not always well understood by researchers:

> The fact that there is little agreement among commentatoɪs writing in statistical books and articles concerning specific courses of action to be followed with multiple comparisons simply means that the issues are complex, and that no single solution can be offered to meet adequately the varied needs of researchers. Consequently, you should view the situation... with a realization that you must work the problem out for yourself.

The purpose of the present paper is to acquaint the reader with some of these complex issues.

Specifically, it is argued that planned comparisons (as against unplanned comparisons and certainly as against omnibus tests involving comparisons across more than two groups) should be employed more frequently in OVA research. And the relative utility of orthogonal (i.e., perfectly uncorrelated) contrasts as against nonorthogonal or correlated comparisons is evaluated. However, prior to presenting these views as three general canons for analytic practice, a context for discussion is established by first explicating three analytic premises.

Three Premises Regarding Analytic Practice

1. **Experimentwise error inflation can be a serious problem, and classical unplanned tests were developed to control inflation of experimentwise error rates.**

Most contemporary researchers recognize that

t-tests performed on all possible pairs of means

> involved in the *F*-test... [to] reveal where significant differences between means lie... is quite unacceptable methodology. The *t*-test was not designed for this use and is invalid when so applied... In spite of the patent invalidity of *t*-testing following a significant *F*-ratio in the analysis of variance, or multiple *t*-testing in lieu of the analysis of variance, this method has often been and continues to be used. (Glass & Stanley, 1970, p. 382)

However, not all researchers understand the basis for these conclusions. The rationale for the conclusions involves the control of experimentwise Type I error rate. A related rationale and the experimentwise error rate problem underlie the use of unplanned comparisons, so the concept of experimentwise error rate merits some discussion.

When a researcher conducts a study in which only one hypothesis is tested, the Type I error probability is the nominal alpha level selected by the researcher, i.e., often the 0.05 level of statistical significance. The probability of making a Type I error when testing a given hypothesis is called the testwise (TW) error rate. Experimentwise (EW) error rate refers to the cumulative probability that one **or more** Type I errors were made **anywhere** in the **full set of all** hypothesis tests conducted in the study. In the case of a study in which only one hypothesis is tested, the testwise error rate exactly equals the experimentwise error rate.

However, when several hypotheses are tested within a single study, the experimentwise error rate may not equal the nominal testwise alpha level used to test each of the separate hypotheses. If all hypotheses are perfectly correlated, then and only then will there be no inflation of experimentwise error rate, because in actuality only one hypothesis is really being tested. If the hypotheses (e.g., the dependent variables) are at all uncorrelated, then there will be at least some inflation of the experimentwise error probability (EW_p). The inflation is at its maximum when the hypotheses are perfectly uncorrelated.

Witte (1985, p. 236) provides an analogy that may clarify why this is so:

> When a fair coin is tossed only once, the probability of heads equals 0.50--just as when a single *t* test is to be conducted at the 0.05 level of significance, the probability of a type I error equals 0.05. When a fair coin is tossed three times, however, heads can appear not only on the first toss but also on the second or third toss, and hence the probability of heads on at least one of the three tosses exceeds 0.50. By the same token, when a type I error can be committed not only on the first test but also on the second or third test, and hence the probability of committing a type I error on at least one of the three tests exceeds 0.05. In fact, the cumulative probability of at least one type I error

can be as large as 0.15 for this series of three _t_ tests.

This coin flip example illustrates a worst-case inflation of experimentwise error (analogized as the flip of a head--H), because the results of each flip are perfectly uncorrelated with previous results (the coin presumably being unaware of or unaffected by its previous behavior). Table 1 illustrates that although the probability of a H on each flip or a fair coin is 50%, the probability of one or more heads over three flips is 87.5%.

INSERT TABLE 1 ABOUT HERE.

In fact, as Thompson (1988c) notes, the experimentwise error rate in a study ranges somewhere between the nominal testwise alpha level (when only one test is conducted or all hypotheses are perfectly correlated) and (1 - (1 - testwise alpha) raised to the power of the number of hypotheses tested (when more than one test is conducted and the hypotheses are perfectly uncorrelated). Love (1988) presents the proof underlying the formula for estimating maximum inflation of experimentwise Type I error. As an example involving estimation of experimentwise error rate, if nine hypotheses were each tested at the 0.05 level in a single study, the experimentwise error rate would range somewhere between 0.05 and 0.37. Table 2 illustrates other calculations of maximum EW error rates for various research situations.

INSERT TABLE 2 ABOUT HERE.

Unplanned comparisons incorporate a correction (Games, 1971a, 1971b) that minimizes the inflation of experimentwise error rate that would otherwise accrue from conducting multiple hypothesis tests in a single study, especially given that omnibus hypotheses have already been tested. As Horvath (1985, p. 223) notes, "Performing a multitude of comparisons between the treatments raises the spectre of an increased overall probability of a Type I error. Post F-test procedures must include some accommodation for this danger." As Kirk (1984, p. 360) explains,

> The principal advantage of this multiple comparison procedure over Student's t is that the probability of erroneously rejecting one or more null hypotheses doesn't increase as a function of the number of hypotheses tested. Regardless of the number of tests performed among p means, this probability remains equal to or less than alpha for the collection of tests.

Snodgrass, Levy-Berger and Haydon (1985, p. 386) note that:

> The post hoc tests for such multiple comparisons all adjust, to one degree or another, for the increase in the probability of a Type I error as the number of comparisons in increased. They differ in the degree to which the probability of a Type I error is reduced.

Various authors discuss which tests are more conservative in this adjustment and which are more liberal. The treatment by Keppel and

Zedeck (1989, pp. 172-180) is especially thoughtful.

2. **Balanced classical factorial OVA and planned orthogonal contrasts <u>both</u> inflate experimentwise error rates to their <u>maximums</u>.**

Experimentwise error rate is at a maximum when the hypotheses tested within an experiment are orthogonal or uncorrelated. For example, the tests of all possible omnibus hypotheses in a factorial multi-way ANOVA (called a "factorial" analysis) with equal numbers of subjects in each cell (called a "balanced" design) are all <u>perfectly</u> uncorrelated. This is why the sums of squares (SOS) for each effect plus the error SOS add up to exactly equal the SOS total. Thus, in a 3x4 ANOVA in which the one two-way omnibus interaction and both main effect omnibus hypotheses are tested at the 0.05 level, the experimentwise error rate would be about 0.14 ($1 - (1 - .05)^3 = 1 - .95^3 = 1 - .8574 = .1426$).

Very few researchers and even fewer textbook authors consciously recognize that inflation of experimentwise occurs in classical OVA methods testing omnibus effects prior to the use of unplanned comparisons. An exception is the textbook written by Glass and Hopkins (1984, p. 374), which acknowledges this dynamic in a footnote. Miller (1966, 1977) also thoroughly explores these issues. The failure to consciously recognize these dynamics can doubtless be traced in some measure to paradigm influences (Thompson, 1989b).

As defined by Gage (1963, p. 95), "Paradigms are models, patterns, or schemata. Paradigms are not the theories; they are rather ways of thinking or patterns for research." Tuthill and

Ashton (1983, p. 7) note that

> A scientific paradigm can be thought of as a socially shared cognitive schema. Just as our cognitive schema provide us, as individuals, with a way of making sense of the world around us, a scientific paradigm provides a group of scientists with a way of collectively making sense of their scientific world.

But scientists usually do not consciously reccgnize the influence of their paradigms. As Lincoln and Guba (1985, pp. 19-20) note:

> If it is difficult for a fish to understand water because it has spent all its life in it, so it is difficult for scientists... to understand what their basic axioms or assumptions might be and what impact those axioms and assumptions have upon everyday thinking and lifestyle.

Even though researchers are usually unaware of paradigm influences, paradigms are nevertheless potent influences in that they tell us what we need to think about, and also the things <u>about which we need not think</u>. As Patton (1975, p. 9) suggests,

> Paradigms are normative, they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of a paradigm that constitutes both its strength <u>and</u> its weaknesses--its strength in that it

> makes action possible; its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradiqm.

Both factorial classical OVA (testing omnibus hypotheses) and planned orthogonal contrasts maximally inflate experimentwise error. More researchers need to become cognizant of both realities. The propensity of many researchers to invariably conduct factorial analyses (and thereby to maximally inflate EW error) is particularly disturbing when researchers test omnibus hypotheses about which they do not care or which they feel they cannot interpret, as perhaps in a five-way omnibus interaction test.

Some researchers always test even omnibus effects that are not of interest because they naively believe that such analyses always increase the probability of detecting statistically significant effects on the omnibus hypotheses that are of interest. This can indeed happen, as illustrated in Table 3. The table first presents results for a hypothetical study in which the researcher is really only interested in the three main effects. For the same data these three tests of interest become statistically significant when the researcher tests the four omnibus interaction hypotheses even though these hypotheses were presumed to not be of interest.

INSERT TABLE 3 ABOUT HERE.

Unfortunately, it is also possible that testing omnibus hypotheses that are not of interest can make effects that are the basis of the research become nonsignificant. Table 4 illustrates how factorial versus nonfactorial analysis of the same data might, for example, yield different conclusions regarding the three main effects in the illustration.

INSERT TABLE 4 ABOUT HERE.

These considerations suggest an important guideline for practice for researchers who overcome paradigm influences:

> I. **Given that testing omnibus hypotheses not of interest can affect the results for the hypotheses actually of substantive interest, and given that maximal inflation of experimentwise error occurs in balanced factorial analysis, test only the model of genuine interest. Do not test omnibus hypotheses that are not of interest or which cannot be interpreted.**

It is an ironic tribute to the power of paradigm influences that the same researchers who consider inflation of experimentwise error rate as one rationale for multivariate statistics (which it is--Fish, 1988) and therefore use multivariate statistics are often somehow blind to the similar inflation of experimentwise error that occurs in classical factorial OVA.

3. **Statistical significance tests are grossly influenced by sample size, and significance considerations should not be primary determinants of analytic choices.**

As is the case for other parametric methods, subsumed as special cases of canonical correlation analysis (Thompson, 1988a), statistical significance tests can be employed to test a null hypothesis that there is zero effect size for a given hypothesis. The propensity to overinterpret significance tests continues, notwithstanding several decades of effort "to exorcise the null hypothesis" (Cronbach, 1975, p. 124). Thompson (1989a, p. 66) notes that

> few statistical procedures have caused more confusion within the research community than statistical significance testing... Because statistical significance is largely an artifact of sample size, significance decisions... must be interpreted in the context of sample size.

Rosnow and Rosenthal (1989, p. 1277) comment on contemporary overemphasis on significance tests:

> It may not be an exaggeration to say that for many PhD students, for whom the .05 alpha has acquired an almost ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation p is less than .05.... [But] surely, God loves the .06 nearly as much as the .05 [level].

Thompson (1987a) explores the consequences of these problems. Even sophisticated authors of prominent textbooks are sometimes not quite sure what role significance tests should play in multivariate analysis (Thompson, 1987b, 1988d), though doctoral students may be disproportionately susceptible to excessive awe for significance tests (Eason & Daniel, 1989; Thompson, 1988b). Recent important treatments of these issues are also offered by Huberty (1987) and by Kupfersmid (1988).

Researchers who have had the fortunate experience of working with large samples (cf. Kaiser, 1976) soon realize that virtually all null hypotheses will be rejected, since "the null hypothesis of no difference is almost never exactly true in the population" (Thompson, 1987a, p. 14). As Meehl (1978, p. 822) notes, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Thus Hays (1981, p. 293) argues that "virtually any study can be made to show significant results if one uses enough subjects."

Presume that a researcher was working in the Houston school district, and analyzed data involving some of the district s 200,000 students. Perchance the researcher decided to compare the mean IQ scores of 12,000 students located in one zip code with the mean IQ of the 188,000 remaining students residing in other zip codes. Since the *t* distribution approaches the *Z* distribution as sample size approaches infinity, researchers use the *Z* distribution to tests mean differences with large samples. These calculations are reported in Table 5.

INSERT TABLE 5 ABOUT HERE.

The mean IQ (100.15, *SD*=15) of the 12,000 students residing in the zip code of interest differs to a statistically significant degree (Zcalc = 2.12 > Zcrit = 1.96, *p*<.05) from the mean (99.85, *SD*=15) of the remaining 188,000 students. The less thoughtful researcher might suggest to school board members that special programs for gifted students should be erected throughout the zip code of the 12,000 students, since they are "significantly" brighter than their compatriots.

Alternatively, the more thoughtful researcher in such a situation would note that the standardized difference in these two means (.3/15 = 0.02) is trivial. The difference of means (.3 = one-third of one IQ point) is also substantially smaller than one standard error of an IQ measure with a reliability coefficient of 0.92, i.e., SEM = SD*((1-r)**.5) = 4.24. Such a thoughtful researcher would be reticent to extrapolate policy recommendations from every statistically significant result.

These considerations suggest that researchers out to interpret results from a canonical analysis by considering significance test results and effect size (Huberty, 1987), or by interpreting significance in the context of sample size (i.e., at what smaller sample size would this result have been no longer significant?--Thompson, 1989a), or by conducting analyses that investigate the replicability of results (Thompson, 1989c). Replicability analyses include the cross-validation logics discussed by Thompson (1984, pp. 41-47, 1989c), or variants of bootstrap (Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1987, in press) or jackknife (e.g., Crask & Perreault, 1977; Daniel, 1989) methods.

Again, it is ironic that researchers who are blinded by the paradigm influences which create an excessive reliance on significance tests are often hoisted on their own petards. The researcher desirous of statistically significant effects for substantive main and interaction effects will quite reasonably employ the largest sample possible so as to achieve the hoped-for results. Regrettably, large samples that tend to yield significance for substantive tests also tend to yield statistically significant results leading to rejection of method assumption null hypotheses, as in the test of equality of dependent variable variances across groups required by the ANOVA homogeneity of variance assumption.

Two Types of Planned Comparisons: Orthogonal versus Nonorthogonal Contrasts

Orthogonal Contrasts Defined

Planned comparisons are the alternative to unplanned comparisons for researchers who wish to isolate differences between sets of specific means. Pedhazur (1982, chapter 9) and Loftus and Loftus (1982, chapter 15) provide valuable explanations of these methods. Various types of planned comparisons can be used, including both orthogonal and non-orthogonal planned comparisons. Planned comparisons typically involve weighting data by sets of "contrasts" such as those presented by Thompson (1985) or the contrasts presented in Table 6. Other types of contrasts, those which test for trends in means, are provided by Fisher and Yates (1957, pp. 90-100) and by Hicks (1973) for various research designs.

INSERT TABLE 6 ABOUT HERE.

Contrasts are typically developed to sum to zero, as do all five contrasts presented in Table 6. Contrasts are uncorrelated or orthogonal (and the hypotheses they represent likewise) when the contrasts each sum to zero and when the cross-products of each pair of contrasts all sum to zero also. It can be demonstrated that these conditions are sufficient to yield perfectly uncorrelated variables. One formula for the Pearson product-moment *r* is:

$$\frac{\mathrm{SUM}_{XY} - ((\mathrm{SUM}_X * \mathrm{SUM}_Y)/N)}{((\mathrm{SUM\ SQUARED\ X'S} - ((\mathrm{SUM}_X)^2/N))(\mathrm{SUM\ SQUARED\ Y'S} - ((\mathrm{SUM}_Y)^2/N)))**.5}$$

Consider only the numerator of the expression. By definition, the sum of the cross-products (SUM_{XY}) is zero. Since by definition both contrasts also sum to zero ($\mathrm{SUM}_X = \mathrm{SUM}_Y = 0$), $\mathrm{SUM}_X * \mathrm{SUM}_Y$ equals zero, and N into zero will also equal zero. Since zero (SUM_{XY}) minus zero ((($\mathrm{SUM}_X * \mathrm{SUM}_Y$)/N)) equals zero, the numerator of the expression is zero. Since any number divided into zero is zero, *r* will be zero, regardless of the divisor.

The contrasts presented in Table 6 are all uncorrelated, based on these requirements. Planned contrasts like those in Table 6 can be employed in a regression analysis in the manner illustrated by Thompson (1985) and as explained by Pedhazur (1982). The required computer cards for this case are presented in Appendix A.

The number of orthogonal planned comparisons always equals the number of degrees of freedom for a given effect. As Hays (1981, p. 425) notes,

> Each and every degree of freedom associated with treatments in any fixed-effects analysis of variance corresponds to some possible comparison of means. The number of degrees of freedom for the mean square between is the number of possible _independent_ [i.e.,

orthogonal] comparisons to be made on the means.

The Case for Orthogonal Contrasts

Most researchers believe that orthogonal planned comparisons have special appeal. Kachigan (1986, p. 310) notes that:

> The importance that we place on a set of orthogonal comparisons is that both of these [individual testwise and experimentwise] significance levels are known to us... On the other hand, when we deal with sets of unplanned non-orthogonal comparisons, these probabilities are not generally available to us, because of the unplanned nature of the comparisons, and because of the non-independence among them.

Keppel (1982, p. 147) suggests that:

> The value of orthogonal comparisons lies in the independence of inferences, which, of course, is a desirable quality to achieve. That is, orthogonal comparisons are such that any decision concerning the null hypothesis representing one comparison is uninfluenced by the decision concerning the null hypothesis representing any other orthogonal comparison. The potential difficulty with nonorthogonal comparisons, then, is interpreting the different outcomes. If we reject the null hypotheses for two nonorthogonal comparisons, which comparison represents the "true" reason for the observed differences?

Of course, orthogonal contrasts applied to balanced designs yield nonoverlapping sums of squares that exactly add to the total sums of squares, just as classical OVA yields uncorrelated or nonoverlapping sum of squares for omnibus tests. It is this "computational simplicity" (Cohen, 1968, p. 440) of orthogonal omnibus tests that led, in part, to the widespread popularity of OVA methods in the era prior to the widespread availability of computers. So another appeal of orthogonal contrasts is that these analyses are analogous in their characteristics to the results in popular omnibus tests.

In summary, using orthogonal contrasts has at least three advantages. First, the exact testwise and experimentwise error rates are both known to us. Second, interpretation tends to be facilitated since equivocal or ambiguous results are less likely. And third, the logic underlying findings can be more readily generalized to the practice in popular omnibus OVA applications using balanced designs, since classical omnibus tests in such cases are also perfectly uncorrelated.

The Case for Nonorthogonal Contrasts

Some researchers do not believe that planned comparisons should necessarily be orthogonal. For example, Winer (1971, p. 175) argues that, "In practice the comparisons that are constructed are those having some meaning in terms of the experimental variables; whether these comparisons are orthogonal or not makes little or no difference." Similarly, even though Cohen and Cohen (1975) called orthogonal "planned comparisons... the most elegant multiple comparison procedure [with] good power characteristics," they noted orthogonal contrasts can "only infrequently be employed in behavioral science investigations because the questions to be put to the data are simply not usually independent" (p. 158).

The primary rationale for using *non*orthogonal contrasts, then, is substantive. Since the number of possible orthogonal contrasts for an effect equals the degrees of freedom for the effect, we may not have available enough orthogonal contrasts to address all the issues of genuine substantive interest. Furthermore, we may be forced by orthogonality constraints to test hypotheses that are not particularly interesting. For example, for a three level way, we may be primarily interested in contrasting the dependent variable mean of level-one subjects against the mean of level-three subjects. Once this first contrast (-1, 0, +1 or +1, 0, -1) is established, to be orthogonal the second contrast (-1, +2, -1 or +1, -2, +1) *must* test whether the dependent variable mean of the level-two subjects differs from the mean of all the subjects in either level one or level three of the way.

Some researchers find these possibilities very troubling. For example, Huberty and Morris (1988, p. 576) argue that, "When a researcher is specifying interesting contrasts, orthogonality need *not* be an issue. One should ask interesting questions, without worrying about redundancy!"

Independent of the fact that experimentwise error is indeterminently inflated when nonorthogonal contrasts are employed, the primary problem with employing numerous nonorthogonal contrasts is that EW error may also be inflated to an unacceptable degree. Although the number of orthogonal contrasts is limited by the degrees of freedom for a given omnibus effect, in some research situations many more nonorthogonal contrasts can be tested for the same effect. The few researchers who have confronted this problem have not yet satisfactorily resolved the issues involved.

Several researchers have suggested using a criterion of reasonableness to decide when EW error inflation is unacceptably high--these researchers have shown a propensity to tolerate idiosyncratic definitions of reasonableness as against seeking a consensus regarding an operational definition of acceptable limits. For example, Huberty and Morris (1988, p. 573, emphasis added) describe the roles they believe significance levels and effect sizes should play in assessing contrast effects:

> Jointly considering the two indicators, p and eta-squared, one can arrive at a conclusion regarding the existence of "real" (i.e., generalizable) contrast effects. Real effects exist if *p* is "small" and if eta-squared is "substantial." The determination of a small *p* value and a substantial eta-squared value is researcher- and situation-dependent. (Of course, the number of contrasts being investigated should be considered.) As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness! There are *no general rules or set criteria for being reasonable*.

Similarly, Miller (1966, p. 35, emphasis added) notes that

> There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his own judgment for the problem at hand. Large single experiments cannot be treated as a whole (family) without an unjustifiable loss in sensitivity.

Unfortunately, subjective definitions of "reasonableness" leave researchers with views of good practice that are not readily commensurable, since no two researchers may agree on their choices regarding what is and what is not reasonable. An approach which emphasizes abstract reasonableness as the standard of practice relies upon the good will and wisdom of the researcher with no basis for determining who has effectively exercised either good will or wisdom. Science does not progress very rapidly absent some agreement regarding epistemology.

A first pass at establishing an acceptable upper limit on inflation of experimentwise Type I error rate inflation might be couched in the context of contemporary practice with classical OVA procedures. Both experimentwise and effectwise limits might be specified.

With respect to experimentwise limits, researchers using balanced factorial OVA designs appear to be willing to tolerate inflation of EW rates equal to the number of possible omnibus tests. For example, for a 4x4x3x2 design, most researchers appear to be willing to tolerate inflation of alpha=.05 to .5367 (1 - (1-

$.05)^{15}$, where 15 = 4 main effects + 6 two-way effects + 4 three-way effects + 1 four-way effect). Thus, under this definition, the researcher would be limited to the use of 15 orthogonal contrasts, unless an additive (p is divided by the number of tests--e.g., .05/15 = .003) or multiplicative Bonferroni correction was invoked (Huberty, 1987).

A more restrictive limit might be set at the effectwise level, in order to be more conservative. The number of orthogonal contrasts for an omnibus effect equals the number of degrees of freedom for the effect. For example, the "A" way in a 4x4x3x2 design has 3 degrees of freedom (4-1), so exactly three orthogonal contrasts are possible. Table 7 presents the contrasts that might be employed for the "A" way. The contrasts "O1", "O2" and "O3" are orthogonal. These three contrasts exactly restate (and only restate) the information contained in the cell information column titled "A". Thus, the multiple correlation (*R*) between "A" and "O1", "O2", and "O3" as a set is exactly 1.0.

INSERT TABLE 7 ABOUT HERE.

As Kirk (1968) notes, most researchers do not adjust for inflation of experimentwise error when they conduct orthogonal planned tests. This suggests a rule that might be applied to decide when to adjust for the EW_p inflation that occurs when nonorthogonal planned tests are utilized: Once the multiple *R* between nonorthogonal contrasts for an omnibus effect and the relevant cell information exceeds 1.0, invoke a Bonferroni correction for EW inflation (Huberty, 1987). This would make the error rate inflation fairly comparable across analytic choices, and would facilitate generalization across the literature.

Table 7 illustrates the possibilities. The contrasts "A1" to "A4" are nonorthogonal tests of differences within the "A" way of the hypothetical 4x4x3x2 design. The multiple correlation between "A1", "A2" and "A3" with the level assignment information ("A") is 1.0. Therefore, a researcher who also wished to test the null hypothesis represented by contrast "A4" (i.e., that the dependent variable mean of the subjects in level four equalled the mean of the subjects in either level one or level two) would invoke the correction. For example, each comparison might be tested at the .0125 (.05/4) level of significance.

These various considerations suggest a second canon regarding analytic practice:

> II. **Given that orthogonal contrasts (a) yield known testwise and experimentwise error rates, (b) tend to yield less ambiguous results, and (c) invoke partitioning of the sum of squares of the dependent variable that is analogous to the orthogonal partitioning invoked in classical OVA, planned orthogonal contrasts should be employed when the contrasts can be used to test the substantive hypotheses of interest. When the use of nonorthogonal contrasts becomes necessary, the researcher should invoke Bonferroni corrections of testwise alpha when the inflation of EW error exceeds the two limits suggested here. In all cases effect size and replicability analyses should be conducted to augment the interpretation of significance tests--such analyses focus on the primary focus in research (generalization), are less starkly influenced by sample size, and place error rate issues in proper perspective.**

The contrast proposed by Huberty and Morris (1988) appears to have special appeal, and warrants further investigation.

Two Reasons Why Planned Comparisons are Superior

There are two reasons why researchers generally prefer the use of planned comparisons to the use of unplanned comparisons (cf. Benton, 1990; Tucker, 1990). First, as noted by numerous researchers, planned comparisons offer more power against making Type II errors:

> procedures recommended for *a priori* orthogonal comparisons are more powerful than procedures recommended for *a priori* nonorthogonal and *a posteriori* comparisons. That is, the former procedures are more likely to detect real differences among means. (Kirk, 1968, p. 95)

> The probability of test's detecting that... [the contrast's effect] is not zero [i.e., is statistically significant] is greater with a planned than with an unplanned comparison on the same sample means. Thus, for any particular comparison, the test is more powerful when planned than when post hoc. (Hays, 1981, p. 438)

> Post hoc tests protect us from making too many Type I errors by requiring a bigger difference before declaring it to be significant than do planned comparisons. But this protection tends to be too

> conservative for planned comparisons, thereby lowering the power of the test. (Minium & Clarke, 1982, p. 322)

> The tests of significance for a priori, or planned, comparisons are more powerful than those for post hoc comparisons. In other words, it is possible for a specific comparison to be not significant when tested by post hoc methods but significant when tested by a priori methods. (Pedhazur, 1982, pp. 304-305 (also Kerlinger & Pedhazur, 1973, p. 131))

> Post hoc comparisons must always follow the finding of a significant overall *F*-value... There are no limits to the number of combinations that can be tested post hoc, but none of these procedures has the power of planned comparison tests for detecting statistical significance. (Sowell & Casey, 1982, p. 119)

> The test of planned subhypotheses is more powerful than the test of post hoc subhypotheses. For this reason, we should make planned comparisons whenever possible in planning the design of research within the ANOVA context. (Glasnapp & Poggio, 1985, p. 474)

Second, and perhaps even more importantly, planned comparisons tend to force the researcher to be more thoughtful in conducting research, since the number of planned comparisons that can be tested is limited. The number of orthogonal planned comparisons cannot exceed the degrees of freedom for an effect, as noted previously. The number of nonorthogonal contrasts would also be limited, if the canons suggested here were accepted.

As Snodgrass, Levy-Berger and Haydon (1985, p. 386) suggest, "The experimenter who carries out post hoc comparisons often has a rather diffuse hypothesis about what the effects of the manipulation should be." Keppel (1982, p. 165) notes that, "Planned comparisons are usually the motivating force behind an experiment. These comparisons are targeted from the start of the investigation and represent an interest in particular combinations of conditions--not in the overall experiment." In summary, as Kerlinger (1986, p. 219) suggests, "While post hoc tests are important in actual research, especially for exploring one's data and for getting leads for future research, the method of planned comparisons is perhaps more important scientifically."

It is important to note that most researchers have fairly good notions of what their studies will show, at least when research is grounded in theoretical constructs or in previous empirical findings, so most researchers are able to suggest planned comparisons prior to data collection. Thus, Huberty and Morris (1988, p. 576) maintain that

> only very few research situations would preclude a researcher from specifying all contrasts of interest

prior to an examination of the outcome measures and/or the outcome 'cell' means. (A typical set of contrasts investigated consists of, simply, all pairwise comparisons.)

A Concrete Heuristic Example of Power

Just as some researchers benefit from seeing heuristic demonstrations that all parametric significance testing procedures are subsumed by and can be conducted with canonical correlation analysis (Thompson, 1988a), it may be helpful to present a hypothetical analysis demonstrating that planned orthogonal comparisons have greater statistical power against Type II error than testing omnibus hypotheses and then exploring significant effects with unplanned comparisons. The data presented in Table 6 can be utilized for this purpose. Table 8 presents a conventional one-way ANOVA keyout associated with the Table 6 data. Even if the researcher conducted unplanned post hoc tests in the absence of a statistically significant main effect, none of the unplanned tests would result in a statistically significant comparison for these data. However, as noted in Table 9, a statistically significant ($p < 0.01$) result is isolated for the hypothesis that the mean attitude-toward-school score of the two school board members differs from the mean for the remaining 10 subjects.

INSERT TABLES 8 AND 9 ABOUT HERE.

The Use of Planned Comparisons in Lieu of Omnibus Tests

Some researchers suggest that at least some unplanned comparisons can be made even if an omnibus effect is not statistically significant. For example, Spence, Cotton, Underwood and Duncan (1983, p. 215) suggest that,

> The Tukey _hsd_ [honestly significant difference test] usually is performed only if the _F_ obtained in the analysis of variance is significant, but it theoretically permissible to perform whatever the significance of _F_.

Similarly, Hays (1981, p. 434) notes:

> This statement is not to be interpreted to mean that post hoc comparisons are somehow illegal or immoral if the original _F_ test is not significant at the required alpha level... What one cannot do is to attach an unequivocal probability statement to such post hoc comparisons, unless the conditions underlying the method have been met.

However, the preponderant view regarding use of unplanned post hoc tests is expressed by Gravetter and Wallnau (1985, p. 423):

> These [a posteriori] tests attempt to control the overall alpha level by making the adjustments for the number of different samples (potential comparisons) in the experiment. To justify a posteriori tests, the _F_-ratio from the overall ANOVA must be significant.

On the other hand, with respect to the use of planned comparisons, "Most statisticians agree that planned _t_ tests between means are appropriate, even when the overall _F_ is insignificant" (Clayton, 1984, p. 193). Snodgrass, Levy-Berger and Haydon (1985, p. 386) concur:

> For planned comparisons, it is not necessary for the overall ANOVA to be significant in order to carry them out... Post hoc comparisons, on the other hand, may not be carried out unless the overall ANOVA is significant.

Gravetter and Wallnau (1985, p. 423) agree that, "Planned comparisons can be made even when the overall _F_-ratio is not significant."

In fact, "It is <u>not necessary</u> to perform an over-all test of significance prior to carrying out planned orthogonal _t_ tests" (Kirk, 1968, p. 73, emphasis added). As Hays (1981, p. 426) suggests,

> The _F_ test gives evidence to let us judge if all of a set of _J_ - 1 such orthogonal comparisons are simultaneously zero in the populations. For this reason, if planned orthogonal comparisons are tested separately, the overall _F_ test is not carried out, and vice versa.

Swaminathan (1989, p. 231, emphasis added) presents the same argument with respect to the MANOVA case:

> The often advocated procedure of following up the rejection of the null hypothesis with a more powerful multiple comparison procedure should be

> discouraged. First, the overall rejection of the null hypothesis does not guarantee any meaningful contrast among the means will be significant, as our example showed. Second..., significant contrasts may be found even when the null hypothesis would not have been rejected. Third, follow up multiple comparison procedures which are unrelated to the overall test result in an inflation of the experiment-wise error rate. If multiple comparisons are of primary interest, a suitable multiple comparison procedure can be used <u>without</u> first performing an overall test.

These considerations suggest a third canon for analytic practice:

> III. **Given that planned tests have greater power against Type II error than either unplanned tests or omnibus tests, planned comparisons should be employed in most research studies using OVA methods. Planned tests should be employed in lieu of omnibus tests.**

Rosnow and Rosenthal (1989, p. 1281) quite rightly deplore the "overreliance on omnibus tests of diffuse hypotheses that although providing protection for some investigators from the dangers of 'data mining' with multiple tests performed as if each were the only one considered" because omnibus tests generally do not:

> tell us anything we really want to know. As Abelson (1962) pointed out long ago in the case of analysis of variance (ANOVA), the problem is that when the null hypothesis is accepted, it is frequently

> because of the insensitive omnibus character of the standard F-test as much as by reason of sizable error variance. All the while that a particular predicted pattern among the means is evident to the naked eye the standard F-test is often insufficiently illuminating to reject the null hypothesis that several means are statistically identical.

Planned contrasts (Rosnow & Rosenthal, 1989, p. 1281) encourage precision of thought and theory, and "usually result in increased power and greater clarity of substantive interpretation."

Summary

The literature regarding the use of multiple comparisons in analysis of variance was reviewed. Three analytic premises provided a framework for the discussion. It was argued that experimentwise Type I error rate inflation can be serious, and that its influences are often unrecognized in many ANOVA applications. It was noted that both classical balanced omnibus and orthogonal planned contrast tests inflate experimentwise error to an identifiable maximum. Finally, it was suggested that significance test results are overinterpreted in contemporary analytic practice, and that researchers must also consider effect sizes and replicability or invariance analyses when formulating interpretations.

Three canons to guide analytic practice were suggested. First, it was suggested that omnibus hypotheses that are not of interest or which cannot be interpreted should not be tested, since such tests can distort the hypothesis tests that are of interest, as illustrated with examples. Second, it is suggested that orthogonal contrasts should be preferred over nonorthogonal contrasts, and that when nonorthogonal contrasts must be used it may be necessary to use a corrected testwise alpha level. Finally, it was suggested that planned contrasts should be used in place of either omnibus or unplanned hypothesis tests. Two reasons why planned comparisons are generally superior were presented. Use of planned comparisons tends to result in more thoughtful research with greater power against Type II error. Throughout the paper small data sets and examples were employed to make the discussion concrete.

References

Benton, R.L. (1990, January). The statistical power of planned comparisons. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED forthcoming)

Clayton, K. N. (1984). An introduction to statistics. Columbus, OH: Merrill.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crask, M.R., & Perreault, W.D., Jr. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Cronbach, L.J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Daniel, L.G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 305 382)

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Eason, S.H., & Daniel, L.G. (1989, January). Trends and methodological practices in several cohorts of dissertations.

Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 306 299)

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. American Psychologist, 29, 25-26.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

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.

Fish, L.J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, England: Oliver and Boyd.

Fisher, R. A., & Yates, F. (1957). Statistical tables for biological, agricultural and medical research (5th ed.). New York: Hafner Publishing Co.

Gage, N.L. (1963). Paradigms for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching (pp. 94-141). Chicago: Rand McNally.

Games, P. A. (1971a). Errata for "Multiple comparisons on means," AERJ, 1971, 531-565. American Educational Research Journal 8, 677-678.

Games, P. A. (1971b). Multiple comparisons of means. American

Educational Research Journal 8, 531-565.

Glasnapp, D. R., & Poggio, J. P. (1985). Essentials of statistical analysis for the behavioral sciences. Columbus, OH: Merrill.

Glass, G. V, & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glass, G. V, & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs: Prentice-Hall.

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

Gravetter, F. J., & Wallnau, L. B. (1985). Statistics for the behavioral sciences. St. Paul, MN. West.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Hicks, C. R. (1973). Fundamental concepts in the design of experiments. New York: Holt, Rinehart and Winston.

Horvath, T. (1985). Basic statistics for behavioral sciences. Boston: Little, Brown and Company.

Huberty, C.J. (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Huberty, C.J, & Morris, J.D. (1988). A single contrast test procedure. Educational and Psychological Measurement, 48, 567-578.

Huck, S. W., Cormier, W. H., & Bounds, Jr., W. G. (1974). Reading statistics and research. New York: Harper & Row.

Kachigan, S. K. (1986). Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods. New York: Radius Press.

Kaiser, H.F. (1976). [Review of Factor analysis as a statistical method]. Educational and Psychological Measurement], 36, 586-589.

Keppel, G. (1982). Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice-Hall.

Keppel, G., & Zedeck, S. (1989). Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches. New York: W.H. Freeman.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole. (pp. 69-98)

Kirk, R. E. (1984). Elementary statistics. Monterey, CA: Brooks/Cole.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Lincoln, Y.S., & Guba, E.G. (1985). Naturalistic inquiry. Beverly Hills: SAGE.

Loftus, G. R., & Loftus, E. F. (1982). Essence of statistics. Monterey, CA: Brooks/Cole.

Love, G. (1988, November). Understanding experimentwise error probability. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Document Reproduction Service No. ED 304 451)

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

Lunneborg, C.E. (in press). [Review of Computer intensive methods for testing hypotheses]. Educational and Psychological Measurement.

McGuigan, F. J. (1983). Experimental psychology: Methods of research (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Miller, R.G. (1966). Simultaneous statistical inference. New York: McGraw-Hill.

Miller, R.G. (1977). Developments in multiple comparisons, 1966-1976. Journal of the American Statistical Association, 72, 779-788.

Minium, E. W., & Clarke, R. B. (1982). Elements of statistical reasoning. New York: John Wiley and sons.

Moore, G. W. (1983). Developing and evaluating educational research. Boston: Little, Brown and Company.

Patton, M.Q. (1975). Alternative evaluation research paradigm. Grand Forks: University of North Dakota Press.

Pedhazur, E. J. (1982). Multiple regression in behavioral research:

Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Snodgrass, J. G., Levy-Berger, G., & Haydon, M. (1985). Human experimental psychology. New York: Oxford University Press.

Sowell, E. J., & Casey, R. J. (1982). Research methods in education. Belmont, CA: Wadsworth.

Spence, J. T., Cotton, J. W., Underwood, B. J., & Duncan, C. P. (1983). Elementary statistics (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Swaminathan, H. (1989) Interpreting the results of multivariate analysis of variance. In B. Thompson (Ed.), Advances in social science methodology (pp. 205-232, Vol. 1). Greenwich, CT: JAI Press.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Beverly Hills: SAGE.

Thompson, B. (1985). Alternate methods for analyzing data from education experiments. Journal of Experimental Education, 54, 50-55.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, 46, 917-928.

Thompson, B. (1987a, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1987b). [Review of Foundations of behavioral research (3rd ed.)]. Educational Research and Measurement, 47, 1175-1181.

Thompson, B. (1988a, April). Canonical correlation analysis: An explanation with comments on correct practice. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1988b, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1988c). Misuse of chi-square contingency table test statistics. Educational and Psychological Research, 8, 39-49.

Thompson, B. (1988d). [Review of Analyzing multivariate data]. Educational and Psychological Measurement, 48, 1129-1135.

Thompson, B.(1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). The place of qualitative methods in contemporary social science: The importance of post-paradigmatic thought. In B. Thompson (Ed.), Advances in social science *methodology* (pp. 1-42, Vol. 1). Greenwich, CT: JAI Press.

Thompson, B. (1989c). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*, 2-6.

Tucker, M.L. (1990, January). *A compendium of textbook views on planned versus post hoc tests*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin. (ERIC Document Reproduction Service No. ED forthcoming)

Tuthill, D., & Ashton, P. (1983). *Improving educational research through the development of educational paradigms*. Educational Researcher, 12(10), 6-14.

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. *Educational Researcher*, *9*, 5-10.

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

Witte, R. S. (1985). *Statistics* d.). New York: Holt, Rinehart and Winston.

Table 1
All Possible Families of Outcomes for a Fair Coin Flipped Three Times

	Flip # 1		2		3
1.	T	:	T	:	T
2.	H	:	T	:	T
3.	T	:	H	:	T
4.	T	:	T	:	H
5.	H	:	H	:	T
6.	H	:	T	:	H
7.	T	:	H	:	H
8.	H	:	H	:	H
p of H on each Flip	50%		50%		50%

p of 1 or more H's (TW error analog) in set of 3 Flips = 7/8 = 87.5%

or

where TW error analog = .50,
EW $p = 1 - (1 - .5)^3$
$= 1-.5^3 = 1-.125 = .875$

Table 2
Maximum Experimentwise Type I Error Inflation

	TW alpha		Tests		Experimentwise alpha
1 - (1 -	0.05)	**	1	=	
1 - (	0.95)	**	1	=	[a]
1 -	0.95			=	0.05000
Range Over Testwise (TW) alpha = .01					
1 - (1 -	0.01)	**	5	=	0.04901
1 - (1 -	0.01)	**	10	=	0.09562
1 - (1 -	0.01)	**	20	=	0.18209
Range Over Testwise (TW) alpha = .05					
1 - (1 -	0.05)	**	5	=	0.22622
1 - (1 -	0.05)	**	10	=	0.40126
1 - (1 -	0.05)	**	20	=	0.64151
Range Over Testwise (TW) alpha = .10					
1 - (1 -	0.10)	**	5	=	0.40951
1 - (1 -	0.10)	**	10	=	0.65132
1 - (1 -	0.10)	**	20	=	0.87842

Note. "**" = "raise to the power of".

[a] These calculations are presented (a) to illustrate the implementation of the formula step by step and (b) to demonstrate that when only one test is conducted, the experimentwise error rate equals the testwise error rate, as should be expected if the formula behaves properly.

Table 3
An Example of How Factorial Analysis can Help Yield Significance for Effects of Interest by Analyzing Even Effects Not of Interest

Nonfactorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**7.4**	**1**	**7.4**	4.144000	4.20	NS
Age	**7.0**	**1**	**7.0**	3.920000	4.20	NS
Sex	**6.0**	**1**	**6.0**	3.360000	4.20	NS
Residual	50.0	28	1.785714			
Total	**70.4**	**31**				

Factorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**7.4**	**1**	**7.4**	5.92	4.26	Rej
Age	**7.0**	**1**	**7.0**	5.6	4.26	Rej
Sex	**6.0**	**1**	**6.0**	4.8	4.26	Rej
2-Way						
Meth*Age	5.0	1	5.0	4.0	4.26	NS
Meth*Sex	5.0	1	5.0	4.0	4.26	NS
Age*Sex	5.0	1	5.0	4.0	4.26	NS
3-Way	5.0	1	5.0	4.0	4.26	NS
Residual	30.0	24	1.25			
Total	**70.4**	**31**				

Note. Entries in **bold** remain constant.

Table 4
An Example of How Factorial Analysis can Hurt by Yielding Nonsignificance for the Effects of Primary Interest

Nonfactorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**8.1**	**1**	**8.1**	4.279245	4.20	Rej
Age	**8.3**	**1**	**8.3**	4.384905	4.20	Rej
Sex	**8.0**	**1**	**8.0**	4.226415	4.20	Rej
Residual	53.0	28	1.892857			
Total	**77.4**	**31**				

Factorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**8.1**	**1**	**8.1**	3.888000	4.26	NS
Age	**8.3**	**1**	**8.3**	3.984000	4.26	NS
Sex	**8.0**	**1**	**8.0**	3.840000	4.26	NS
2-Way						
Meth*Age	(.5	1	0.5	0.240000	4.26	NS
Meth*Sex	1.0	1	1.0	0.480000	4.26	NS
Age*Sex	1.0	1	1.0	0.480000	4.26	NS
3-Way	0.5	1	0.5	0.240000	4.26	NS
Residual	50.0	24	2.083333			
Total	**77.4**	**31**				

Note. Entries in **bold** remain constant.

Table 5
Test of Mean Differences for School District Example

```
Z = (  M1    -  M2   ) / (((SD1**2/   n1 ) + (SD2**2/   n2  )) ** .5)
Z = (100.15 - 99.85) / (((15**2 / 12000) + (15**2 / 188000)) ** .5)
  =       0.3        / ((( 225  / 12000) + ( 225  / 188000)) ** .5)
  =       0.3        /  ((   0.01875     +     0.001196    )  ** .5)
  =       0.3        /   (           0.019946808             ** .5)
  =       0.3         /     0.141233170
  =            2.124146887
```

Table 6
Hypothetical Data for Attitudes Toward School Study (n=12)

				Contrast				
Group	LEVEL	ID	DV	C1	C2	C3	C4	C5
Students	1	1	10	-1	-1	-1	-1	-1
		2	20	-1	-1	-1	-1	-1
Teacher Aides	2	3	10	1	-1	-1	-1	-1
		4	20	1	-1	-1	-1	-1
Teachers	3	5	10	0	2	-1	-1	-1
		6	20	0	2	-1	-1	-1
Principals	4	7	10	0	0	3	-1	-1
		8	20	0	0	3	-1	-1
Superintendents	5	9	10	0	0	0	4	-1
		10	20	0	0	0	4	-1
Board Members	6	11	25	0	0	0	0	5
		12	35	0	0	0	0	5

Table 7
Various Contrasts Used to Predict Cell Assignment in Way "A"

A	O1	O2	O3	A1	A2	A3	A4
1	-1	-1	-1	-1	-1	-1	-1
2	1	-1	-1	1	0	-1	-1
3	0	2	-1	0	-1	2	0
4	0	0	3	0	0	0	2

Note. $R_{AX(O1,O2,O3)}$ = 1.0. $R_{AX(A1,A2,A3)}$ = 1.0.

Table 8
One-Way ANOVA Results

Source	SOS	df	Mean Square	F	p	eta Square
Between	375.0000	5	75.0000	1.5000	.3155	.55556
Error	300.0000	6	50.0000			
Total	675.0000	11				

Table 9
Planned Comparison Results

Contrast Source	SOS	df	Mean Square	F	p	eta Square
C1	.0000	1	.0000	0.0000		.00000
C2	.0000	1	.0000	0.0000		.00000
C3	.0000	1	.0000	0.0000		.00000
C4	.0000	1	.0000	0.0000		.00000
C5	375.0000	1	375.0000	12.5000	.0054	.55556
Error	300.0000	6	50.0000			
Total	675.0000	11				

APPENDIX A
Selected SPSS-X Control Cards

```
TITLE '*****OMNIBUS no  POSTHOC no  A PRIORI yes'
FILE HANDLE BT/NAME='APRIORI.DTA'
DATA LIST FILE=BT/LEV 1 DV 2-4
COMPUTE C1=0
COMPUTE C2=0
COMPUTE C3=0
COMPUTE C4=0
COMPUTE C5=0
IF (LEV EQ 2)C1=1
IF (LEV EQ 1)C1=-1
IF (LEV EQ 3)C2=2
IF (LEV EQ 1 OR LEV EQ 2)C2=-1
IF (LEV EQ 4)C3=3
IF (LEV LT 4)C3=-1
IF (LEV EQ 5)C4=4
IF (LEV LT 5)C4=-1
IF (LEV EQ 6)C5=5
IF (C5 EQ 0)C5=-1
REGRESSION VARIABLES=DV C1 TO C5/DESCRIPTIVES=ALL/
  CRITERIA=PIN(.95) POUT(.999) TOLERANCE(.00001)/DEPENDENT=DV/
  ENTER C5/ENTER C4/ENTER C3/ENTER C2/ENTER C1/
```