ADVANCES IN SOCIAL SCIENCE METHODOLOGY

Editor: BRUCE THOMPSON

Volume 3 • 1994

ADVANCES IN SOCIAL SCIENCE METHODOLOGY

Editor: BRUCE THOMPSON
Texas A&M University
and
Baylor College of Medicine

VOLUME 3 • 1994

JAI PRESS INC.

Greenwich, Connecticut *London, England*

55 Old Post Road, No. 2
Greenwich, Connecticut 06830

JAI PRESS LTD.
The Courtyard
28 High Street
Hampton Hill
Middlesex TW12 1PD
England

ISBN: 1-55938-379-8

Manufactured in the United States of America

ADVANCES IN SOCIAL SCIENCE METHODOLOGY

Volume 3 • 1994

LIST OF CONTRIBUTORS

Elisa B. Benitez	Department of Educational Psychology University of Georgia
J. Douglas Carroll	Graduate School of Management Rutgers University
Tungshan F. Chou	Department of Educational Psychology University of Georgia
Allen L. Edwards	Department of Psychology University of Washington
Lynne K. Edwards	Department of Psychology University of Minnesota
Carl J Huberty	Department of Educational Psychology University of Georgia
Roger E. Kirk	Department of Psychology Baylor University
Margaret D. LeCompte	College of Education University of Colorado
Stanley A. Mulaik	School of Psychology Georgia Institute of Technology
Judith Preissle	Department of Social Science Education University of Georgia
Janet C. Rice	School of Public Health Tulane University
Bruce Thompson	Department of Educational Psychology Texas A&M University
Sharon L. Weinberg	Program of Quantitative Studies, SEHNAP New York University

FOREWORD

The hallmark of this series has always been the clarity with which the authors have presented syntheses of previous work or recent advances impacting contemporary analytic practice. This volume, Volume 3 in the series, is no different.

One difference, however, is that Volume 3 includes a section of four chapters focusing on a single analytic issue: the use of multiple comparisons or contrasts. The topic of multiple contrasts has generated considerable controversy. These authors address the issues in a comprehensive manner. The reader will be well situated in understanding these issues after reading these four chapters.

The remaining four chapters focus on developments involving a variety of methodology choices. These range from qualitative research, multidimensional scaling, logistic regression, to philosophy of science as related to the use of multivariate statistics. Again in this section the series authors present lucid remarks and thoughtful insights.

Bruce Thompson
Series Editor

CONTENTS

PART I

THE USE OF MULTIPLE CONTRASTS

PLANNED VERSUS UNPLANNED AND ORTHOGONAL VERSUS NONORTHOGONAL CONTRASTS:

THE NEOCLASSICAL PERSPECTIVE

Bruce Thompson

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, 1991). Most users of ANOVA-type methods (ANOVA, ANCOVA, MANOVA, MANCOVA—hereafter labeled 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, p. 68). 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 4×3 design the omnibus test for the

Advances in Social Science Methodology, Volume 3, pages 3–27.

ISBN: 1-55938-379-8

four-level A way has 4 – 1 or 3 degrees of freedom). Gravetter and Wallnau concur: "Reject H_0 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" (1985, p. 423). Moore 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. (1983, p. 299)

Contrasts or *comparisons* can be used to test more specific hypotheses about particular differences in means. A contrast is a coding vector that actually represents a given hypothesis. Contrasts are usually developed such that the numbers constituting each contrast sum to zero. Table 1 presents examples for a study involving a one-way six-level design, and let us say, two subjects per level.

For example, contrast C1 in Table 1 tests the null hypothesis that the dependent variable mean of the two subjects in level 1 of the way equals the dependent variable mean of the two subjects in level 2 of the way, ignoring the eight subjects in the remaining cells of the design. Contrasts can be thought of as being applied to cell means, or to the data for each subject (as illustrated for this design later in the chapter). There are many types of contrasts, including those that are planned as against unplanned, and those that are uncorrelated or orthogonal as against nonorthogonal.

Thus, *unplanned* (also called *a posteriori* or *post hoc* or *unfocused*) multiple comparison tests (e.g., Scheffe, Tukey, or Duncan) are among the choices that can be used to isolate means that are significantly different within OVA ways (also called factors) having more than two levels. As Glass and Hopkins 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" (1984, p. 368).

Textbook authors tend to discuss unplanned comparison or contrast procedures in somewhat pejorative terms. For example, Kirk speaks of the use of unplanned

Table 1. Example Contrasts for a One-Way Six-Level Design

	Contrasts				
Level	*C1*	*C2*	*C3*	*C4*	*C5*
1	–1	–1	–1	–1	–1
2	1	–1	–1	–1	–1
3	0	2	–1	–1	–1
4	0	0	3	–1	–1
5	0	0	0	4	–1
6	0	0	0	0	5

comparisons as "ferreting out significant differences among means, or, as it is often called, data snooping" (1984, p. 360). 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 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 commentators 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. (1982, p. 164)

The purpose of the present chapter 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

Premise 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, that is, often the .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. Of course, in the case of a study in which only one hypothesis is tested, the TW error rate exactly equals the EW error rate.

However, when several hypotheses are tested within a single study, the EW error rate may not equal the nominal TW 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 EW 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 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. (1985, p. 236)

This coin flip example illustrates a worst-case inflation of EW 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 2 illustrates that although the probability of a H on each flip of a fair coin is 50%, the probability of one or more Hs over three flips is 87.5%.

In fact, as Thompson (1988c) explains, the EW error rate in a study ranges somewhere between the nominal TW 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 EW Type I error. As an example involving estimation of EW error rate, if nine hypotheses were each tested at the .05 level in a single study, the experimentwise error rate would range somewhere between .05 and .37. Table 3 illustrates other calculations of maximum EW error rates for various research situations.

Unplanned comparisons incorporate a correction (Games, 1971a, 1971b) that minimizes the inflation of EW error rate that would otherwise accrue from conducting multiple hypothesis tests in a single study, especially given that omnibus

Table 2. 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%

Note. *Probability of 1 *or more* Hs (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 3. Maximum EW Type I Error Inflation

TW alpha	*Tests*	*EW alpha*
1 – (1 – 0.05) **	1 =	
1 – (0.95) **	1 =	
1 – 0.95	=	0.05000[a]
	Range over 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 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 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 EW error rate equals the TW error rate, as should be expected if the formula behaves properly.

hypotheses have already been tested. As Horvath 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" (1985, p. 223). As Kirk 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. (1984, p. 360)

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

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.

Premise 2. *Balanced classical factorial OVA and planned orthogonal contrasts* both *inflate EW error rates to their* maximums.

EW 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 multiway ANOVA (called a *factorial* analysis) with equal numbers of subjects in each cell (called a *balanced* design) are all *perfectly* 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 3×4 ANOVA in which the one two-way omnibus interaction and both main effect omnibus hypotheses are tested at the .05 level, the EW error rate would be about .14 $[1 - (1 - .05)^3 = 1 - .95^3 = 1 - .8574 = .1426]$.

Very few researchers and even fewer textbook authors consciously recognize that inflation of EW error rates occurs in classical OVA methods testing omnibus effects prior to the use of unplanned comparisons. One 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, "Paradigms are models, patterns, or schemata. Paradigms are not the theories; they are rather ways of thinking or patterns for research" (1963, p. 95). Tuthill and Ashton note that

> A scientific paradigm can be thought of as a socially shared cognitive schema. Just as our cognitive schema provides 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. (1983, p. 7)

But scientists usually do not consciously recognize the influence of their paradigms. As Lincoln and Guba 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. (1985, pp. 19–20)

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 *about which we need not think.* As Patton 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 *and* 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 paradigm. (1975, p. 9)