DOCUMENT RESUME

ED 306 244 TM 013 013

AUTHOR	Thompson, Bruce
TITLE	Five Steps for Improving Evaluation Reports by Using Different Data Analysis Methods.
PUB DATE	Mar 89
NOTE	49p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Analysis of Covariance; Analysis of Variance; *Data Analysis; Decision Making; Educational Research; *Evaluation Methods; Evaluation Problems; Evaluation Utilization; Multivariate Analysis; *Program Evaluation; Regression (Statistics); *Research Methodology; *Statistical Analysis; Statistical Significance
IDENTIFIERS	*Evaluation Reports

ABSTRACT

Although methodological integrity is not the sole determinant of the value of a program evaluation, decision-makers do have a right, at a minimum, to be able to expect competent work from evaluators. This paper explores five areas where evaluators might improve methodological practices. First, evaluation reports should reflect the limited contribution that statistical significance testing can make to the interpretation of results. Second, evaluation reports should also reflect the fact that multivariate statistics are often vital in educational research because multivariate methods best honor the reality to which the researcher is trying to generalize. Third, evaluation reports should reflect the recognition that discarding variance to conduct chi-square or analyses of variance and covariance and multivariate analyses of variance and covariance (OVA methods) can lead to serious distortions in interpretations. Even when OVA methods are appropriate, the methods should usually be implemented using regression approaches. Fourth, evaluation reports should reflect the recognition that covariance statistical correlations are usually least helpful when corrections are most needed. Finally, evaluation reports should reflect the recognition that stepwise analytic methods can lead to seriously distorted interpretations. Fourteen tables and three figures illustrate these contentions. A 102-item list of references is provided. (SLD)

himprove.rno 3/10/89

FIVE STEPS FOR IMPROVING EVALUATION REPORTS

BY USING DIFFERENT DATA ANALYSIS METHODS

Bruce Thompson

University of New Orleans 70148

Paper present_d at the annual meeting of the American Educational Research Association (session #6.45), San Francisco, March 27, 1989.

ABSTRACT

Methodological integrity is not the *sina qua non* of program evaluation endeavors. But decision-makers do have a right to expect, at a minimum, that evaluation results will generalize to reality. Five precepts regarding improved methodological practice are presented and explained in some detail. These precepts focus on statistical significance testing, the use of multivariate statistics to honor the complexity of program reality, the discarding of variance to conduct OVA or chi- quare analyses, the use of covariance or statistical control, and the use of stepwise analytic methods.

Program evaluators have frequently been concerned that decision makers do not use their evaluation results to formulate policy (King & Thompson, 1983b). These concerns have a long history. Wholey, Scanlon, Duffy, Fukumoto, and Vogt (1970, p. 46) concluded that "the recent literature is unanimous in announcing the general failure of evaluation to affect decision-making in a significant way." Weiss (1972, p. 319) wrote that "evaluators complain about many things, but their most common complaint is that their findings are ignored." Similarly, Rippey (1973) concluded that

> At the moment, there seems to be no evidence that evaluation, although the law of the land, contributes anything to educational practice other than headaches for the researcher, threats for innovators, and depressing articles for journals devoted to evaluation. (p. 9)

Prominent evaluation theorists have expressed these concerns. In 1973 Worthen and Sanders noted that "evaluation is one of the most widely discussed but little used processes in today's educational systems." Stake (1976, p. 1) once wrote, "We do not know whether or not evaluation is going to contribute more to the problems of education or more to the solutions." Two years later Patton (1978) wrote that

> In many ways the odds are all against utilization and it is quite possible to become skeptical about the futility of trying to have impact in a world where situation after situation seems impervious to change. (p. 291)

Alkin and Daillak (1979, p. 41) noted that "there have been great hopes for evaluation, not only among evaluators themselves, but also among other educators, elected officials, and the public. Yet these hopes have dimmed."

Views expressed by local education agency (LEA) practitioners can be especially dramatic. For example, Holley (1980) noted that

> In an ideal world we wouldn't have to worry about utilization. Educators would be eagerly awaiting our findings and would promptly rush to put them into practice. I don't need to tell you that isn't happening. (p. 2)

Kilbourne and DeGracie (1979) note that

> All LEAs, with possibly a few exceptions, can point to their volumes of research and evaluation verbiage sitting on the shelves of district administrators being used for little else than a door stop, swatting flies, or any other of the various and sundry purposes for which research is used in the publis schools. (p. 12)

More recently, researchers and evaluation practitioners (Thompson, 1982b) have recognized that there are many types of use, some of which are subtle and somewhat difficult to discern in the real world of incremental decision making (Eason, 1988; King, Thompson & Pechman, 1981). Indeed, such a view does help explain the seemingly paradoxical finding (Alkin, Kosecoff, Fitz-Gibbon & Seligman, 1974; King & Thompson, 1983a) that admin:strators consistently report evaluation to be usefu˙. And administrators and evaluators appear to have reasonably similar views of types of evaluators and of the evaluation process (Thompson, 1980; Thompson & Miller, 1984).

There have been numerous empirical studies of evaluation, summarized by King and Thompson (1983b), and by King et al. (1981) in an even more comprehensive report. A number of studies of evaluation use have employed simulated evaluation reports (e.g., Thompson, 1982a). Although such studies can be criticized on several grounds (Thompson & King, 1981), these reports have provided some insight into evaluation dynamics (Thompson & Levitov, 1983).

Both simulation and other studies (e.g., naturalistic qualitative studies) and experience all suggest that "the personal factor" (Cronbach et al., 1980, p. 174; Guskin, 1980, p. 21; Holley, 1980, p. 8; Leviton & Hughes, 1979, p. 21), i.e., administrator perception that the evaluator is competent and trustworthy, is critical to making evaluation use occur. In some senses evaluation is largely a persuasive endeavor (Eason, 1988; Thompson, 1981). This suggests that the methodology employed in evaluation repcrts has a relatively minor role to play in influencing use.

Patton et al. (1977, p. 151) report that "there is little in our data to suggest that improving methodological quality in and of itself will have effect on increasing the utilization of evaluation research." Similarly, Dickey (1980) found that methodological sophistication had no relationship to the level of use. Kennedy, Apling and Neumann (1980) conducted interviews and found that evaluators worried about the methodological quality of their work, but that report users were more concerned about "educaticnal quality" (p. 111) than about technical quality.

All this is not to say that methodological quality is unimportant. Rather, it is argued that methodological issues must be kept in perspective with respect to the impacts that technical merit will have. As Alkin (1975, p. 207) notes, "one of the best conscientious defenses against non-utilization of evaluation findings is a technically sound, methodologically credible study." As Leviton and Hughes (1979, p. 25) suggest, "If [methodological] quality does influence use, it is likely to do so primarily through increased trust that the findings are an accurate picture of the program." As Johnson (1978, p. 12) notes, "The central message in this regard is that it is not enough to conduct methodologically sound research."

But decision makers do have a right, at a minimum, to be able to expect technically competent work from evaluators. Such work does not guarantee use, but may be a necessary requisite to possible use. The purpose of the present paper is to explore five areas where evaluators might improve methodological practice. Some evaluators already adhere to these methodological precepts. Methodology is in some ways an ideologically driven business (Cliff, 1987; Thompson, 1988d), but the practice of some evaluators might be improved with reflection upon the issues raised herein.

A Preliminary Caveat: The Role cf Statistical Method in Inquiry

However, one preliminary caveat is in order--methodological integrity is not the ultimate sina qua non of research, evaluation or otherwise. Certainly it is true that, "Although the quality of educational research is improving, evidence still indicates that much of the research published has important weaknesses" (Borg, 1983, p. 193). Empirical studies of methodological practice in published research confirm these general impressions (Persell, 1976; Wandt, 1965; Ward, Hall & Schramm, 1975). Some of the problems in the quality of the research literature can be attributed to the journal review process, studied in an intriguing fashion by Peters and Ceci (1982). Nevertheless, as Glass (1979, p. 12) suggests, "Our research literature in education is not of the highest quality, but I suspect that it is good enough on most topics."

1. Evaluation reports should reflect the limited contribution that statistical significance testing can be make to the interpretation of results.

Few methodological offerings have sparked more controversy than Sir Ronald Fisher's promulgation of significance testing methods, methods that apparently were developed prior to Fisher's work (Carlson, 1976). The past 30 years have involved periodic efforts "to exorcise the null hypothesis" (Cronbach, 1975, p. 124). Morrison and Henkel (1970) and Carver (1978) provide historically important and incisive explanations of the limits of significance testing as an aid to interpretation. More recent informative treatments are available from Dar (1987), Huberty (1987), Kupfersmid (1988), and Thompson (1987b, 1988c, in press-b).

Most researchers have been taught the statistical significance of results does not inform the researcher regarding the *importance* of outcomes. Shaver (1985, p. 58) makes this point in a concrete fashion in his contrived dialogue about significance testing:

> Chris: [Looking puzzled.] Well, as I said, it [my result] was statistically significant. You know, that means it wasn't likely to be just a chance occurrence... An unlikely occurrence like that *surely* must be important.
>
> Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?

Yet, in three ways actual behavior tends to belie a failure to really accept that significance testing does not inform decisions regarding the importance of results. First, journal editorial boards tend to perceive articles that report significant results more favorably than articles not reporting significant results (Atkinson, Furlong & Wampold, 1982). Second, readers of research findings tend to perceive more favorably those articles reporting statistically significant results (Cohen, 1979). Third, and most disturbing of all, authors tend not to submit manuscripts in which nonsignificant results must be reported, and even tend to abandon lines of inquiry on the basis of such results (Greenwald, 1975). These behaviors have been too readily transmitted to evaluators.

Too few researchers and evaluators appreciate which study features contribute to statistical significance. Although significance is a function of at least seven interrelated features of a study (Schneider & Darcy, 1984), sample size is the primary influence on significance. Some example results may clarify the ways in which sample sizes affect significance tests.

Tables 1 and 2 present significance tests associated with varying sample sizes and either moderate (9.8%) or larger (33.6%) fixed effect sizes, respectively. The tables can be viewed as presenting results for either a multiple regression analysis involving two predictor variables (in which case the "r sq" effect size would be called the squared multiple correlation coefficient, R) or an analysis of variance involving an omnibus test of differences in three means in a one-way design (in which case the "r sq" effect size would be called the correlation ratio or eta squared).

INSERT TABLES 1 AND 2 HERE.

Each table presents results for fixed effect sizes but increasing sample sizes (4, 13, 23, 33, 43, 53, 63, or 123). For the fixed effect size of 9.8% involved in Table 1, the fixed effect size becomes statistically significant when there are somewhere between 53 and 63 subjects in the analysis. For the 33.6% effect size reported in Table 2, the result becomes statistically significant when there are somewhere between 13 and 23 subjects in the analysis.

For a fixed effect size, adding subjects to the analysis impacts statistical significance in two ways. First, as illustrated in Tables 1 and 2, the critical F at a fixed alpha gets smaller as degrees of freedom error increase. Second, as the degrees of freedom error increase, the mean square error gets smaller, and thus the calculated F gets larger.

The eva' ıator who does not genuinely understand statistical significance would differentially interpret the effect size of 9.8% when there were 53 versus 63 subjects, and would differentially interpret the fixed effect size of 33.6% when there were 13 versus 23 subjects in the analysis. Yet the effect sizes within each table are fixed. Empirical studies of research practice indicate that superficial understanding of significance testing has actually led to serious distortions such as researchers interpreting significant results involving small effect sizes while ignoring nonsignificant results involving large effect sizes (Craig, Eison & Metze, 1976)!

Nor does significance testing typically inform the evaluator regarding the likelihood that results will be replicated in future work (Carver, 1978). Evaluators who wish to estimate the likely replicabi''ty of results should instead employ cross-validation logic (Campo, 1988), the "jackknife" logic developed by Tukey and his colleagues (Crask & Perreault, 1977), or the "bootstrap" logic developed by Efron and his associates (Diaconis & Efron, 1983).

Two aspects of significance testing interpretation in evaluation reports warrant attention. First, some evaluators use language implying that they are interpreting significance tests as if they were effect sizes. But, as Kerlinger (1986, p. 214) emphasizes, "Tests of statistical significance like *t* and *F* unfortunately do not indicate the magnitude or strength of relations." Yet Kerlinger (1986) himself constantly refers to results being "highly significant" (cf. pp. 187, 248, 334), and other respected textbook authors do so as well (e.g., Cliff, 1987, p. 394).

A second problem in language, implying the interpretation of significance tests as effect sizes, involves the use of phrases such as "the results approached statistical significance." Robert Brown, former editor of the *Journal of College Student Personnel*, made the humorous but telling comment at a recent conference: "How do these authors know their results weren't trying to avoid statistical significance?" Yet evaluators too often find themselves reporting that, "The number of years of experience was not significant but did approach significance."

The most serious misinterpretations of significance testing tend to occur when sample size is small and effect sizes are large but are underinterpreted, or when sample sizes are commendably large and are statistically significant but effect sizes are modest and are overinterpreted. Evaluators can conduct analyses such as those reported to in Tables 1 and 2 to determine when statistically significant effects become insignificant as sample size decreases, or when an effect size becomes statistically significant as sample size is increased. Such analyses give the administrator a valuable perspective for interpreting results. The importance of such analyses has been recognized in Guidelines for Authors in some journals:

> 6. Authors are encouraged to assist readers in interpreting statistical significance of their results. For example, results may be indexed to sample size. An author may wish to say, "this correlation coefficient would have been statistically significant even if sample size had been as small as _n_=33," or "this correlation coefficient would have been statistically significant if sample size had been increased to _n_=138." (_MECD_, 1988, p. 46)

2. _Evaluation reports should reflect the fact that multivariate statistics are often vital in educational research_.

Multivariate statistics have been available to researchers for many years, although even today "there are many articles in the research literature in which multiple univariate statistics are calculated rather than a single multivariate analysis; for instance, one article may report 50 _t_-tests rather than one MANOVA" (Moore, 1983, p. 307). McMillan and Schumacher (1984) isolated one reason why some researchers have hesitated to use multivariate statistical methods:

> The statistical procedures for analyzing many variables at the same time have been available for many years, but it has only been since the computer age that researchers have been able to utilize these

> procedures. There is thus lag in training of researchers that has militated against the use of these more sophisticated procedures. There are in evidence more each year in journals, however... (p. 270)

Hinkle, Wiersma and Jurs (1979) concurred, noting that "it is becoming increasingly important for behavioral scientists to understand multivariate procedures even if they do not use them in their own research." And recent empirical studies of research practice do confirm that multivariate methods are employed with some regularity in published behavioral research (Elmore & Woehl..e, 1988; Gaither & Glorfeld, 1985; Goodwin & Goodwin, 1985).

There are two reasons why multivariate methods are so important in behavioral research, as noted by Thompson (1986b) and by Fish (1988). First, multivariate methods control the inflation of Type I "experimentwise" error rates. Most researchers are familiar with "testwise" alpha. But while "testwise" alpha refers to the probability of making a Type I error for a given hypothesis test, "experimentwise" error rate refers to the probability of having made a Type I error anywhere within the study. When only one hypothesis is tested for a given group of people in a study, "experimentwise" error rate will exactly equal the "testwise" error rate.

But when more than one hypothesis is tested in a given study, the two error rates will not be equal. Witte (1985, p. 236) explains the two error rates using an intuitively appealing example involving a coin toss. If the toss of heads is equated with a Type I error, and if a coin is tossed only once, then the probability of a head on the one toss and of at least one head within the set of one toss will both equal 50%. But if the coin is tossed three times, even though the "testwise" probability of a head on each given toss in 50%, the "experimentwise" probability that there will be at least one head in the whole set of three flips will be inflated to more than 50%. Researchers control "testwise" error rate by picking small values, usually 0.05, for the "testwise" alpha. "Experimentwise" error rate, on the other hand, can be controlled at the "testwise" level by employing multivariate statistics.

When researchers test several hypotheses in a given study, but do not use multivariate statistics, the "experimentwise" error rate will range somewhere between the "testwise" error rate and the ceiling calculated in the manner illustrated in Table 3. Where the experimentwise error rate will actually lie will depend upon the degree of correlation among the dependent variables in the study. Because the exact rate in a practical sense is readily estimated only when the dependent variables are perfectly correlated (and "experimentwise" error will equal the "testwise" error) or are perfectly uncorrelated (and "experimentwise" error will equal the ceiling calculated in the manner illustrated in Table 3), it is particularly disturbing that the researcher may not even be able to determine the exact "experimentwise" error rate in some studies!

INSERT TABLE 3 ABOUT HERE.

Paradoxically, although the use of several univariate tests in a single study can lead to too many hypotheses being spuriously rejected, as reflected in inflation of "experimentwise" error rate, it is also possible that the failure to employ multivariate methods can lead to a failure to identify statistically significant results which actually exist. Fish (1988) provides a data set illustrating this equally disturbing possibility. The basis for this paradox is beyond the scope of the present treatment, but involves the second major reason why multivariate statistics are so important.

Multivariate methods are often vital in behavioral research because multivariate methods best honor the reality to which the researcher is purportedly trying to generalize. Since significance testing and error rates may not be the most important aspect of research practice (Thompson, 1988c), this second reason for employing multivariate statistics is actually the more important of the two grounds for using these methods. Thompson (1986b, p. 9) notes that the reality about which most researchers wish to generalize is usually one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects." As Hopkins (1980, p. 374) has emphasized:

> These multivariate methods allow understanding of relationships among several variables not possible with univariate analysis... Factor analysis, canonical correlation, and discriminant analysis--and modifications of each procedure--allow researchers to study complex data, particularly

> situations with many interrelated variables. Such is the case with questions based in the education of human beings.

Similarly, McMillan and Schumacher (1984) argue that:

> Social scientists have realized for many years that human behavior can be understood only by examining many variables at the same time, not by dealing with one variable in one study, another variable in a second study, and so forth... These [univariate] procedures haved failed to reflect our current emphasis on the multiplicity of factors in human behavior... In the reality of complex social situations the researcher needs to examine many variables simultaneously. (pp. 269-270)

3. <u>Evaluation reports should reflect the recognition that discarding variance to conduct chi-square or OVA analyses can lead to serious distortions in interpretations, and that even when OVA methods are appropriate the methods should usually be implemented using regression approaches</u>.

Cohen (1968, p. 441) has characterized the conversion of intervally scaled variables down to the nominal level of scale as the "squandering [of] much information." As Kerlinger (1986, p. 558) explains, this squandering can lead to distorted results:

> ...Partitioning a continuous variable into a dichotomy or trichotomy throws information away... To reduce a set of values with a relatively wide range to a dichotomy is to reduce its variance and thus its possible correlation with other variables.

Thompson (1988a, pp. 3-4) notes that

> Variance is the "stuff" of which all quantitative research studies are made... It is not usually sensible to invest serious effort in collecting reliable and valid continuous score data, and to then casually discard the information that we previously went to some trouble to collect.

Evaluators too frequently discard variance in order to conduct either Pearson chi-square contingency table tests or ANOVA, ANCOVA, MANOVA or MANCOVA (hereafter labelled OVA methods). Certainly there are many problems with typical applications of the chi-square contingency table test (Thompson, 1988b), but OVA methods are more frequently applied (Elmore & Woehlke, 1988; Gaither & Glorfeld, 1985; Goodwin & Goodwin, 1985), and empirical research indicates that the use of OVA methods with variables that were originally intervally scaled does introduce distortions (Thompson, 1986a). Thus, Cliff (1987, p. 130) correctly criticizes the practice of discarding variance on intervally scaled predictor variables to perform OVA analyses:

> Such divisions are not infallible; think of the persons near the borders. Some who should be highs are actually classified as lows, and vice versa. In addition, the "barely highs" are classified the same as the "very highs," even though they are different. Therefore, reducing a reliable variable to a dichotomy makes the variable more unreliable, not less.

Furthermore, even when intervally scaled variables are naturally nominally scaled, regression approaches to OVA analyses still tend to be superior to classical OVA calculations (Thompson, 1985).

Most evaluators employing OVA methods are aware that "A researcher cannot stop his analysis after getting a significant _F_" (Huck, Cormier & Bounds, 1974, p. 68). 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."

Many evaluators employ unplanned (also called _a posteriori_ or post hoc) multiple comparison tests (e.g., Sheffe, Tukey, or Duncan) to isolate which means are significantly different within OVA ways (also called factors) having more than two levels. Textbook authors tend to discuss unplanned comparisons in somewhat prejorative terms. For example, several authors refer to the application of these comparisons as "data snooping" (Kirk, 1968, p. 73, 1984, p. 360; Pedhazur, 1982, p. 305). Keppel (1982, p. 150) makes reference to "milking" in his discussion of these tests. Similarly, Minium and Clarke (1982, p. 321) note that:

> 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.

Planned (also called a priori or focused) comparisons provide a valuable alternative to unplanned comparisons. Pedhazur (1982, chapter 9) and Loftus and Loftus (1982, chapter 15) provide readable explanations of these comparisons. Planned comparisons typically involve weighting data by sets of "contrasts" such as those presented by Thompson (1985) or those presented in Table 4. 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).

INSERT TABLE 4 ABOUT HERE.

Contrasts are typically developed to sum to zero, as do all five contrasts presented for the data in Table 4. The data represent a hypothetical evaluation study conducted to determine whether various clinical groups score differently on a psychological measure. Contrasts are uncorrelated or orthogonal (as are the hypotheses they represent or test) when the contrasts each sum to zero and when the sum of the cross-products of each pair of contrasts all sum to zero also. Thus, the contrasts presented in Table 4 are uncorrelated.

Some theorists do not believe that planned comparisons should necessarily be orthogonal. For example, Winer (1971, p. 175) argues that "whether these comparisons are orthogonal or not makes little or no difference." However, orthogonal planned comparisons do have special appeal, for statistical reasons delineated elsewhere (Kachigan, 1986, p. 309). But as Keppel (1982, p. 147) suggests:

> 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 regarding 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?

There are two reasons why planned comparisons are usually superior to unplanned comparisons. First, as noted by numerous researchers (Glasnapp & Poggio, 1985, p. 474; Hays, 1981, p. 438; Kirk, 1968, p. 95; Minium & Clarke, 1982, p. 322; Pedhazur, 1982, pp. 304-305; Sowell & Casey, 1982, p. 119), planned comparisons offer more power against Type II errors than do unplanned comparisons, for reasons explained elsewhere (Games, 1971a, 1971b). For example, for the data presented in Table 4, the omnibus test of differences among the six group means is not statistically significant (F=1.5, df=5/6, p=.3155). Furthermore, even if unplanned comparisons were conducted in violation of conventional practice (since the omnibus test was not statistically significant), statistically significant differences would not have been identified either. However, a planned comparison involving the mean of the two level-six subjects versus the mean of the remaining 10 subjects would have been statistically significant (*F*=12.5, *df*=1/6, *p*=.0054).

However, significance is not the end-all and be-all of evaluation research (Thompson, 1988c). The more important reason why planned comparisons are important is that *planned comparisons tend to force the evaluator to be more thoughtful in conducting research*, since planned comparisons must be carefully formulated before data are collected and since typically only a limited number of planned comparisons can be stated in a given study. 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." As Keppel (1982, p. 165) notes,

> 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.

Thus, 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."

4. *Evaluation reports should reflect a recognition that covariance statistical corrections are usually least helpful (and are most dangerous) when corrections are most needed.*

Many "statistical controls" can be invoked to adjust posttest scores when the evaluator believes that or random assignment or design selection have failed to create groups that were equivalent at the start of the experiment or quasiexperiment. These statistical controls are available throughout the entire gamut of quantitative methods. For example, Gorsuch (1983, pp. 89-90) notes that the first factor extracted in a factor analysis can be located to pass directly through a "covariate" variable in factor space. Since factors are uncorrelated, the effects of the first factor on all other factors will have been statistically controlled.

Though many of these statistical controls date back to the beginning of the century (Nunnally, 1975, p. 9), most of the controls have not enjoyed wide use. Analysis of covariance (ANCOVA), for example, has been used in about four percent of the recently published research (Goodwin & Goodwin, 1985, pp. 8-9; Willson, 1980, p. 7). As explained by McGuigan (1983, p. 230):

> Briefly this technique enables you to obtain a measure of what you think is a particularly relevant extraneous variable that you are not controlling. This usually involves some characteristics of your participants. For instance, if you are conducting a study of the effect of certain psychological variables on weight, you might use as your measure the weight of your participants before you administer your experimental treatments. Through analysis of covariance, you then can "statistically control" this variable--that is, you can remove the effect of initial weight from your dependent variable scores, thus decreasing your error variance.

One problem with statistical controls is that they assume very reliable measurement of the control variables. For example, Nunnally (1975, p. 10) notes that reliability will not usually have an appreciable influence on the substantive interpretation of most statistical procedures as long as reliability of measurement is at least 0.70, but that "Measurement reliability becomes crucial... in employing statistical partialling operations, as in the analysis of covariance or in the use of partial correlational analysis." Cliff (1987, p. 129) concurs, noting that

> In general, partial correlation analysis is affected by any lack of reliability or validity in the variables. In many ways these effects resemble tuberculosis as it occurred a generation or two ago: They are widespread, the consequences are serious, the symptoms are easily overlooked, and most people are unaware of their etiology or treatment.

Unfortunately, too many evaluators may not consider and certainly dc not report the measurement error of their variables. As Willson (1980, p. 9) comments, "That reliability of instruments is unreported in almost half the published research is likewise inexcusable at this late date."

Statistical control has been particularly appealing to some evaluators when random assignment was not performed. These researchers expect the statistical adjustments of ANCOVA to magically make groups equivalent.

However, the primary difficulty with statistical control performed to make groups equivalent involves the homogeneity or regression assumption of the methods. The methods assume that the relationship between the covariate and the dependent variable is equivalent in all experimental groups. This assumption is necessary because the statistical control procedures are implemented by adjusting the dependent variable to the extent that the covariate and the dependent variable are correlated *when group membership information is completely ignored*.

Campbell and Erlebacher (1975) present a concrete illustration of how the use of statistical controls can seriously distort evaluation findings when the homogeneity of regression assumption is not met. ANCOVA has been very appealing in research investigating the effects of compensatory education programs. In these cases the treatment intervention is made available to all or most children who are eligible. The control group usually consists of children who were not eligible for the treatment and, therefore, the group is inherently different in its character than the treatment group. In these analyses both the dependent variable and the covariate are cognitive variables. The statistical control procedure assumes that the relationship between the two variables is the same in both groups, i.e., since correlation is a measure of the slope of the regression line for the two variables, that children who are eligible for and receive compensatory interventions *learn at the same rate* as children who are not eligible for the intervention.

The decision to blithely use the statistical control when the homogeneity of regression assumption is not met leads to "tragically misleading analyses" that actually "can mistakenly make compensatory education look harmful" (Campbell & Erlebacher, 1975, p. 597). Similarly, Cliff (1987, p. 273) argues that, "It could be that the <u>relationship</u> between the dependent variable and the covariate is different under different treatments. Such occurrences tend to invalidate the interpretation of the simple partial correlations described above."

Persons who wish to use statistical controls of this type are usually trapped in a nasty dilemma. If the controls are not needed then they should not be used. But if statistical control is needed because the groups in a study are not equivalent, then often the homegeneity of regression assumption cannot be met and the use results in seriously distorted inferences.

It is interesting to note that many evaluators do not recognize the paradox of testing both analytic assumptions and substantive hypotheses for statistical significance. Evaluators frequently try to obtain as large a sample as possible, so that chances for "significance" of substantive tests are maximized. This practice also leads to greater likelihood that tests of homogeneity of variance or of regression will also be significant.

The fallacious use of statistical control in inappropriate ANCOVA applications needs to be recognized by more evaluators, as some theoreticians have long warned of these various dangers (Elashoff, 1969; Lord, 1960). ANCOVA is a special case of regression analysis. As Cliff (1987, p. 275) notes, "We could say that we are fitting a single regression equation to the data for all the groups and then doing an anova of the deviation from the regression line."

Consider the hypothetical data presented in Table 5. The hypothetical study involves four children from a compensatory program ("A") who have lower mean achievement (-.19) on the cognitive pretest ("ZX") than do their peers (mean=.19) from the noncompensatory group. Furthermore, as one might expect, and as illustrated in Figure 1 (which also presents the cognitive posttest ("ZY") scores of the eight children), the children in the two groups are learning at different rates.

INSERT TABLE 5 AND FIGURE 1 ABOUT HERE.

Nevertheless, the ANCOVA procedures employs the single beta weight (_r_ = beta weight for two variable case = .81) derived by ignoring the group membership ("A" or "B") of the children, i.e., derived by ignoring the fact that the children are learning at different rates. This beta weight adjustment is presented in Figure 1 as the regression line for the variables, derived ignoring group membership. However, Figure 1 also indicates that the slopes of regression lines computed separately for the two groups are different, and that it is not reasonable to use the same adjustment for both groups.

Table 6 presents conventional ANOVA results for this data set when no covariance adjustments are implemented. Table 7 presents an ANCOVA utilizing pretest scores ("ZX") as a covariate. Table 8 presents an ANOVA performed on the residual raw scores ("YE" = "ZY" - "YHAT"); this analysis demonstrates that ANCOVA is an ANOVA on posttest scores once the posttest scores have been residualized with the covariate ("YE") in a regression analysis completely ignoring group membership information.

INSEPT TABLES 6 THROUGH 8 ABOUT HERE.

What many evaluators do not understand is how ANCOVA can make the experimental intervention appear less effective. Figure 2 represents a case in which the covariate ("X") is associated with the dependent variable ("Y"), but not with the assignment to experimental conditions ("A"). In other words, the homogeneity of regression assumption is met.

Table 9 presents a one-way ANOVA corresponding to the Figure 2 Venn diagram. Table 10 presents the related ANCOVA. In this example all the adjustment involving the covariate involves variance in the dependent variable not associated with assignment to experimental conditions. Therefore, the sum of squares for the main effect remains unchanged, but the covariate does reduce the sum of squares for error. This results in a smaller mean square error, and thereby a larger calculated _F_ for the main effect.

INSERT FIGURE 2 AND TABLES 9 AND 10 ABOUT HERE.

But Figure 3 presents a case where the homogeneity of regression assumption is not met. Tables 11 and 12 present the related ANOVA and ANCOVA results, respectively. Although the intervention does has some effect, the application of the covariate in this "worst case" example makes the intervention appear entirely ineffective. Clearly, covariance adjustments can have effects that some researchers do not recognize.

INSERT FIGURE 3 AND TABLES 11 AND 12 ABOUT HERE.

The fact that ANCOVA is simply ANOVA on the residual raw scores may also be disturbing from an interpretation point of view. The evaluator took a variable that presumably had some meaning ("ZY"), made an adjustment on it, and was left with an analysis of a residual raw score that, unlike the original dependent variable, has little intrinsic meaning. The result might be difficult to interpret even if the adjustment was reasonable, i.e., if the homogeneity of regression assumption had been met.

Too many researchers and evaluators blindly apply ANCOVA absent an understanding or either the method's logic or its pivotal assumptions. As McGuigan (1983, p. 231) has observed, ANCOVA

> can be seriously misused, and one cannot be assured that it can "save" a shoddy experiment. Some researchers overuse this method as in the instance of a person I once overheard asking of a researcher, "Where is your analysis of covariance?"--the understanding in his department was that it is <u>always</u> used in experimentation.

Of course, the preceeding discussion of the ANCOVA case generalizes to the various types of statistical control that are available to researchers.

ANCOVA is <u>not</u> robust to the violation of the homogeneity of regression assumption, but some evaluators routinely decline to evaluate this assumption. It is ironic that this ineptness may have contributed to the rise of alternative paradigms for conducting evaluation use research (Thompson, in press-a). The failure to test the homogeneity of regression assumption can lead to serious misinterpretations of program effects.

5. Evaluation reports should reflect the recognition that stepwise analytic methods can lead to seriously distorted interpretations.

Stepwise analytic methods may be among the most popular research practices employed in both substantive and validity research. As commonly employed, these methods allow the entry of predictor variables one step at a time, and at each step the removal of previously entered variables is also considered. The methods seem to be somewhat casually employed especially in regression and discriminant analysis research, though variants are also available when other techniques are used (cf. Thompson, 1984, pp. 47-51).

With respect to regression applications, Marascuilo and Serlin (1988, p. 671) note that, "The most popular method in use for selecting the fewest number of predictor variables necessary to guarantee adequate prediction is based on a model referred to as *stepwise regression*." Huberty (in press) concurs, suggesting that "The conduct of analytical procedures in 'steps' is quite common... [These] procedures have enjoyed widespread use by social and behavioral researchers." Unfortunately, stepwise methods can lead to serious misinterpretations of results, and "social science research is replete with misinterpretations of this kind" (Pedhazur, 1982, p. 168).

Three problems with stepwise methods merit special emphasis. First, many evaluators, thanks to "canned" computer programs, do not employ the correct degrees of freedom when evaluating changes in explained variance, i e., usually changes in squared _R_ or lambda. For example, in a stepwise regression analysis, the evaluator at step two may add a second predictor variable into a prediction equation. The evaluator might test the significance of the change in squared _R_ by an _F_ test using 1 and _n_-_q_-1 degrees of freedom, where _q_ is the number of predictor variables in the last step. The numerator degrees of freedom reflects a premise that only one additional predictor variable was employed to yield the squared _R_ change, but _ignores_ the fact that the added predictor was selected by consulting empirical sample results involving a larger set of candidates for entry into the prediction process. Thus, the process ignores that fact that, "in a sense, _all_ the variables are in the equation, even though some of them have [effectively] been given zero weights" (Cliff, 1987, p. 187). Consequently, Cliff (1987, p. 185) suggests that "most computer programs for [stepwise] multiple regression are positively satanic in their temptation toward Type I errors."

Second, some evaluators incorrectly interpret stepwise results in which _q_ predictor variables have been selected as indicating that the predictor variables are the best variables to use if the predictor variable set is limited to size _q_. In fact, in a stepwise analysis in which three steps are conducted, and predictors _A_, _B_, and _C_ are employed, it is entirely possible that three different predictors would represent the optimal predictor set of size three. Stepwise methods select the next-best predictor at each step, given the presence of previous predictors--this is not the same as selecting the optimal predictor variable set of size _q_. As Huberty (in press) notes, "It is generally understood by methodologists that the first _q_ variables entered into either a regression analysis or a discriminant analysis do not necessarily constitute the 'best' subset of size _q_."

Third, some evaluators incorrectly consult order of entry information to evaluate the importance of various predictor variables. As Huberty (in press) explains,

> The first variable entered with a stepwise regression analysis is determined by the correlation between each predictor variable and the criterion variable... The third, say, variable to be entered (and often considered to be the third most important) is dependent on the two variables already entered. If one or two the variables already entered would be changed, then the third variable entered may also be different. This dependence or conditionality truly makes variable importance as determined by stepwise analyses _very_ questionable.

The small data set for a population (_N_=12) presented in Table 13 can be employed to illustrate how sampling error can seriously distort the interpretation of stepwise results involved in predicting dependent variable _ZY_. Table 14 indicates that the three predictor variables share little variance with each other and that the order of predictor variable explanatory power is, respectively, ZX1, ZX2, ZX3, and ZX4.

INSERT TABLES 13 AND 14 ABOUT HERE.

Presume that the evaluator draws a random sample of nine subjects from the population of 12 persons. Each of 55 random collections of nine subjects (omit subjects 1,2,3; omit 1,2,4; etc.) is equally probable. For these illustrative data, only eight samples (omit 1,2,5; 1,2,7; 2,3,7; 2,3,10; 3,4,5; 5,6,8; 7,8,9; and 7,8,12) enter the four predictor variables in the order that is known to be correct when the true population parameters are consulted.

Indeed, only 23 samples select predictor ZX1 as the first prediction entry. Sixteen samples select ZX2 as the first entry; 10 samples select ZX3 as the first variable entered; six samples select the worst predictor, ZX4, as the first or best predictor of ZY. For the sample omitting subjects 3, 4 and 9, the predictor variables are entered in the order: ZX4, ZX2, ZX3, and ZX1.

Clearly, sampling error can seriously distort stepwise results. As Kachigan (1986, p. 265) argues,

> there is the danger that we might select variables for inclusion in the regression equation based on chance relationships. Therefore, as stressed in our discussion of multiple correlation, we should apply our chosen regression equation to a fresh sample of objects to see how well it does in fact predict values on the criterion variable. This validation procedure is absolutely essential if we

are to have any faith at all in the future applicati›ns of the regression equation.

Alternatively, the evaluator might employ a cross-validation procedure such as the one recommended by Huck, Cormier and Bounds (1974, p. 159).

Given these considerations, Kerlinger (1986, p. 545) argues that "the research problem and the theory behind the problem [and not stepwise methods] should determine the order of entry of variables in multiple regression analysis." Evaluators who choose to employ stepwise methods, particularly if they also fail to use replication or cross-validation methods, might best consider Cliff's (1987, pp. 120-121) argument that "a large proportion of the published results using this method probably present conclusions that are not supported by the data."

Summary

As noted previously, methodological integrity is not the sina qua non of program evaluation endeavors. But decision-makers do have a right to expect, at a minimum, that evaluation results will generalize to reality. Five precepts regarding improved methodological practice were presented and explained in some detail. These precepts focus on statistical significance testing, the use of multivariate statistics to honor the complexity of program reality, the discarding of variance to conduct OVA or chi-square analyses, the use of covariance or statistical control, and the use of stepwise analytic methods.

References

Alkin, M.C. (1975). Evaluation: Who needs is? Who cares? *Studies in educational evaluation*, *1*, 201-212.

Alkin, M.C., & Daillak, R.H. (1979). A study of evaluation utilization. *Educational evaluation and policy analysis*, *1*, 41-49.

Alkin, M.C., Kosecoff, J.B., Fitz-Gibbon, C., & Seligman, R. (1974). *Evaluation and decision-making: The Title VII experience.* (CSE monograph series in evaluation No. 4). Los Angeles: UCLA Center for the Study of Evaluation.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189-194.

Borg, W.R. (1983). *Educational research: An introduction* (4th ed.). New York: Longman.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 1), (pp. 597-617). Beverly Hills: SAGE.

Campo, S.F. (1988, November). *Alternative logics for estimating whether results will generalize*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY.

Carlson, R. (1976). Discussion: The logic of tests of significance. *Philosophy of Science*, *43*, 116-128.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.

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

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *47*, 421-423.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and omega-squared. *Bulletin of the Psychonomic Society*, *7*, 280-282.

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.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psycholigists. *American Psychologist, 42*, 145-151.

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

Dickey, B. (1980). Utilization of evaluations of small scale educational projects. *Educational evaluation and policy analysis, 2*, 65-77.

Eason, S.H. (1988). *The effect of persuasion on the utilization of program evaluation information*. Unpublished master's thesis, University of New Orleans, 1988. [Available through interlibrary loan: UNO Interlibrary Loan Office; University of New Orleans; New Orleans, LA 70148]

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6*, 383-401.

Elmore, P.B., & Woehlke, P.L. (1988, April). *Research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

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

Gaither, N., & Glorfeld, L. (1985). An evaluation of the use of tests of significance in organizational behavior research. *Academy of Management Review, 10*, 787-793.

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 on 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. (1979). Policy for the unpredictable (uncertainty research and policy). *Educational Researcher*, *8*(9), 12-14.

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.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Guskin, A.E. (1980). Knowledge utilization and power in university decision making. In L.A. Braskamp & R.D. Brown (Eds.), *Utilization of evaluative information*. San Francisco: Jossey-Bass.

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.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1979). *Applied statistics for the behavioral sciences*. Chicago: Rand McNally.

Holley, F. (1980, April). *Catch a falling star: Promoting the utilization of research and evaluation findings*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hopkins, C.D. (1980). *Understanding educational research: An inquiry approach*. New York: Holt, Rinehart and Winston.

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

Huberty, C. J. (in press). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 1). Greenwich, CT: JAI Press.

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

Johnson, M.D. (1978, April). Evaluation as a model for decision-oriented research. Paper presented at the annual meeting of the Institutional Research Annual Forum. (ERIC Document Reproduction Service No. ED 154 743)

Kachigan, S. K. (1986). Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods (2nd ed.). New York: Radius Press.

Kennedy, M.M., Apling, R., & Neumann, W.F. (1980). The role of evaluation and test information in public schools. Cambridge, MA: The Huron Institute.

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

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

Kilbourne, R., & DeGracie, J. (1979, April). The use of L.E.A. research at the local level: The picture of a dropout. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

King, J.A., & Thompson, B. (1983a). How principals, superintendents view program evaluation. NASSP Bulletin, 67, 46-52.

King, J.A., & Thompson, B. (1983b). Research on school use of program evaluation: A literature review and research agenda. Studies in Educational Evaluation, 9, 5-21.

King, J., Thompson, B., & Pechman, E.M. (1981). Improving evaluation use in local school settings. (Final report on grant NIE-G-80-0082). New Orleans: Orleans Parish Public Schools. (ERIC Document Reproduction Service No. ED 214 998; Ms. #2556 from Psychological Documents--Select Press; P.O. Box 37; Corte Madera, CA 94925)

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

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.

Leviton, L.C., & Hughes, E.F.X. (1979). Utilization of evaluations: A review and synthesis. Evanston, IL: Center for Health Services and Policy Research, Northwestern University.

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

Lord, F. M. (1960). Large sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 309-321.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: W. H. Freeman.

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

McMillan, J.H., & Schumacher, S. (1984). *Research in education: A conceptual approach*. Boston: Little, Brown.

MECD. (1988). Guidelines for authors. *Measurement and Evaluation in Counseling in Development*, *21*, 46.

Minium, W. 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.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Nunnally, J. C. (1975). Psychometric theory--25 years ago and now. *Educational Researcher*, *4*(10), 7-14, 19-20.

Patton, M.Q. (1978). *Utilization-focused evaluation*. Menlo Park, CA: SAGE.

Patton, M.Q., Grimes, P.S., Gutherie, K.M., Brennan, N.J., Grench, B.D., & Blyth, D.A. (1977). In search of impact: An analysis of utilization of federal health evaluation research. In C.H. Weiss (Ed.), *Using social research in public policymaking*. Lexington, MA: Heath.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Persell, C.H. (1976). *Quality, careers and training in educational and social research*. Bayside, NY: General Hall.

Peters, D.P., & Ceci, S.J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences*, *5*, 187-255.

Rippey, R.M. (1973). *Studies in transactional evaluation*. Berkeley: McCutchan.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation*

Review, 8, 573-582.

Shaver, J.P. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

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.

Stake, R. (1976). Evaluation design, instrumentation, data collection and analysis of data. Urbana, IL: Center for Instructional Research and Curriculum Evaluation.

Thompson, B. (1980). Validity of an evaluator typology. Educational Evaluation and Policy Analysis, 2, 59-65.

Thompson, B. (1981, October). Communication theory as a framework for evaluation use research: Evaluation as persuasion. Paper presented at the annual meeting of the Evaluation Network Society, Austin. (ERIC Document Reproduction Service No. ED 209 330)

Thompson, B. (1982a). Administrators' perceptions of various evaluation report styles. Texas Tech Journal of Education, 9, 17-21.

Thompson, B. (Ed.). (1982b, March). Evaluation issues from an LEA perspective. Proceedings from a symposium presented at the annual meeting of the American Educational Research Association, New York. (ERIC Document Reproduction Service ED 216 046)

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Menlo Park: SAGE.

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

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

Thompson, B. (1986b, November). Two reasons why multivariate methods are usually vital. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.

Thompson, B. (1987, 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 Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287

868)

Thompson, B. (1988a). Discarding variance: A cardinal sin in research. *Measurement and Evaluation in Counseling and Development*, *21*, 3-4.

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

Thompson, B. (1988c). A note about significance testing. *Measurement and Evaluation in Counseling and Development*, *20*, 146-147.

Thompson, B. (1988d). Review of *Analyzing multivariate data* by N. Cliff. *Educational and Psychological Measurement*, *48*, 1129-1135.

Thompson, B. (in press-a). The place of qualitative research in contemporary social science. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1). Greenwich, CT: JAI Press.

Thompson, B. (in press-b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*(1).

Thompson, B., & King, J.A. (1981). A critique of evaluation-use research methods. *CEDR Quarterly*, *14*, 19-21. Reprinted in *Louisiana Education Research Journal*, 1982, *7*, 1-6.

Thompson, B., & Levitov, J.E. (1983). Quantitative analysis of previous research on use of evaluation. *Psychological Reports*, *53*, 215-221.

Thompson, B., & Miller, L.A. (1984). Administrators' and evaluators' perceptions of evaluation. *Educational and Psychological Research*, *4*, 207-219.

Wandt, E. (Ed.). (1965). *A cross-section of educational research*. New York: David McKay.

Ward, N.E.W., Hall, B.W., & Schramm, C.F. (1975). *American Educational Research Journal*, *12*, 109-128.

Weiss, C.H. (1972). Utilization of evaluation: Toward comparative study. In C.H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education*. Boston: Allyn & Bacon.

Wholey, J.S., Scanlon, J., Duffy, H., Fukumoto, J., & Vogt, L. (1970). *Federal evaluation policy: Analyzing the effects of public programs*. Washington, DC: The Urban Institute.

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher*, *9*(6), 5-10.

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

Witte, R.S. (1985). Statistics (2nd ed.). New York: Holt, Rinehart and Winston.

Worthen, B.R., & Sanders, J.R. (1973). Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones.

Table 1
Statistical Significance at Various Sample Sizes
for a Fixed Effect Size (Moderate Effect Size)

Source	SOS	r sq	df	MS	F calc	F crit	Decision
SOSexp	98.7	0.098473	2	49.35	0.054614	200.00	Not Rej
SOSunexp	903.6		1	903.6			
SOStot	1002.3		3	334.1			
SOSexp	98.7	0.098473	2	49.35	0.546148	4.10	Not Rej
SOSunexp	903.6		10	90.36			
SOStot	1002.3		12	83.525			
SOSexp	98.7	0.098473	2	49.35	1.092297	3.49	Not Rej
SOSunexp	903.6		20	45.18			
SOStot	1002.3		22	45.55909			
SOSexp	98.7	0.098473	2	49.35	1.638446	3.32	Not Rej
SOSunexp	903.6		30	30.12			
SOStot	1002.3		32	31.32187			
SOSexp	98.7	0.098473	2	49.35	2.184594	3.23	Not Rej
SOSunexp	903.6		40	22.59			
SOStot	1002.3		42	23.86428			
SOSexp	98.7	0.098473	2	49.35	2.730743	c3.19	Not Rej
SOSunexp	903.6		50	18.072			
SOStot	1002.3		52	19.275			
SOSexp	98.7	0.098473	2	49.35	3.276892	3.15	Rej
SOSunexp	903.6		60	15.06			
SOStot	1002.3		62	16.16612			
SOSexp	98.7	0.098473	2	49.35	6.553784	3.07	Rej
SOSunexp	903.6		120	7.53			
SOStot	1002.3		122	8.215573			

Table 2
Statistical Significance at Various Sample Sizes
for a Fixed Effect Size (Larger Effect Size)

Source	SOS	r sq	df	MS	F calc	F crit	Decision
SOSexp	337.2	0.336426	2	168.6	0.253495	200.00	Not Rej
SOSunexp	665.1		1	665.1			
SOStot	1002.3		3	334.1			
SOSexp	337.2	0.336426	2	168.6	2.534957	4.10	Not Rej
SOSunexp	665.1		10	66.51			
SOStot	1002.3		12	83.525			
SOSexp	337.2	0.336426	2	168.6	5.069914	3.49	Rej
SOSunexp	665.1		20	33.255			
SOStot	1002.3		22	45.55909			
SOSexp	337.2	0.336426	2	168.6	7.604871	3.32	Rej
SOSunexp	665.1		30	22.17			
SOStot	1002.3		32	31.32187			
SOSexp	337.2	0.336426	2	168.6	10.13982	3.23	Rej
SOSunexp	665.1		40	16.6275			
SOStot	1002.3		42	23.86428			
SOSexp	337.2	0.336426	2	168.6	12.67478	c3.19	Rej
SOSunexp	665.1		50	13.302			
SOStot	1002.3		52	19.275			
SOSexp	337.2	0.336426	2	168.6	15.20974	3.15	Rej
SOSunexp	665.1		60	11.085			
SOStot	1002.3		62	16.16612			
SOSexp	337.2	0.336426	2	168.6	30.41948	3.07	Rej
SOSunexp	665.1		120	5.5425			
SOStot	1002.3		122	8.215573			

Table 3
"Testwise" and "Experimentwise" Error Rates for Selected Studies

"Testwise"		"Experimentwise" Rate			
Rate	Minimum			n of Tests	Maximum
05.0%	05.0%	1 - (- 05.0%)	**	1	=
05.0%	05.0%	1 - (95.0%)	**	1	=
05.0%	05.0%	1 - 95.0%			= 05.00%
05.0%	05.0%	1 - (- 05.0%)	**	5	= 22.62%
05.0%	05.0%	1 - (- 05.0%)	**	10	= 40.13%
05.0%	05.0%	1 - (- 05.0%)	**	20	= 64.15%

Note. An alpha of 0.05 equals an alpha of 05.0%. "**" means "raised to the power of". The first several rows of the table illustrate the that "testwise" and "experimentwise" error rates are the same when only one test is conducted.

Table 4
Hypothetical Validity Study Data

Group	ID	DV	C1	C2	C3	C4	C5
1	1	10	-1	-1	-1	-1	-1
	2	20	-1	-1	-1	-1	-1
2	3	10	1	-1	-1	-1	-1
	4	20	1	-1	-1	-1	-1
3	5	10	0	2	-1	-1	-1
	6	20	0	2	-1	-1	-1
4	7	10	0	0	3	-1	-1
	8	20	0	0	3	-1	-1
5	9	10	0	0	0	4	-1
	10	20	0	0	0	4	-1
6	11	25	0	0	0	0	5
	12	35	0	0	0	0	5

Table 5
Hypothetical ANCOVA Data Set

Group	ZY	ZX	ZYZX	YHAT	YE
A	-.88	-1.68	1.48	-1.36	.48
A	-.44	-.68	.30	-.56	.11
A	.00	.31	.00	.25	-.25
A	.44	1.30	.57	1.06	-.62
B	-1.32	-.68	.90	-.56	-.77
B	-.44	-.19	.08	-.15	-.29
B	.88	.56	.49	.45	.43
B	1.76	1.06	1.86	.86	.91

Note. The beta weight for the covariance procedure (.813) equals the sum of the cross products (ZXZY) of ZX and ZY divided by _n_-1 (5.694/n-1). The predicted posttest score (YHAT) is each child's pretest (ZX) multiplied by the beta weight. The error in each prediction (YE) is equal to ZY minus YHAT.

Table 6
Conventional ANOVA Results

Source	Sum of Squares	_df_	Mean Squares	_F_	Effect Size
Treatment	.39	1	.39	.35	.056
"Error"	6.61	5	1.10		
Total	7.00	7	1.00		

Note. Effect size is a _r_ squared analog.

Table 7
ANCOVA Results

Source	Sum of Squares	*df*	Mean Squares	*F*	Effect Size
Covariate	4.63	1	4.63	9.95	.661
Treatment	.04	1	.04	.08	.006
"Error"	2.33	5	.47		
Total	7.00	7	1.00		

Note. Effect size is a *r* squared analog.

Table 8
ANOVA Results Using YE as Dependent Variable

Source	Sum of Squares	*df*	Mean Squares	*F*	Effect Size
Treatment	.04	1	.04	.08	.006
"Error"	2.33	5	.47		
Total	2.37	6			

Table 9
ANOVA Associated with Figure 2

Source	Sum of Squares	df	Mean Squares	F Calc	F Crit
Treatment	35	1	35.00	4.85	5.12
"Error"	65	9	7.22		
Total	100	10			

Table 10
ANCOVA Associated with Figure 2

Source	Sum of Squares	df	Mean Squares	F Calc	F Crit
Covariate	20	1	20.0		
Treatment	35	1	35.00	6.22	5.32
"Error"	45	8	5.62		
Total	100	10			

Table 11
ANOVA Associated with Figure 3

Source	Sum of Squares	df	Mean Squares	F Calc	F Crit
Treatment	20	1	20.00	2.25	5.12
"Error"	80	9	8.89		
Total	100	10			

Table 12
ANCOVA Associated with Figure 3

Source	Sum of Squares	df	Mean Squares	F Calc	F Crit
Covariate	30	1	30.0		
Treatment	0	1	.00	.00	5.32
"Error"	70	8	8.75		
Total	100	10			

Table 13
Standardized Data for Five Variables

ID	ZY	ZX1	ZX2	ZX3	ZX4
1	.790	1.422	.350	.322	-.313
2	-1.589	.112	-1.239	-1.094	-.365
3	.127	-.965	.271	.201	-.060
4	-1.656	-2.167	-.498	-.970	.218
5	.176	-1.291	.153	2.393	.159
6	-.017	.636	-1.607	-.168	-1.746
7	-.397	-.173	.931	-.112	-1.704
8	-.594	.532	-.108	.092	.127
9	.846	.528	1.237	-.092	.035
10	.810	.642	-1.400	1.135	1.654
11	1.764	.373	1.290	-.543	1.005
12	-.260	.352	.620	-1.163	.989

Table 14
Bivariate Correlation Matrix

	ZY	ZX1	ZX2	ZX3	ZX4
ZY		.497	.444	.384	.319
ZX1	24.7%		.018	-.074	-.004
ZX2	19.7%	.0%		-.054	.099
ZX3	14.7%	.5%	.3%		.103
ZX4	10.2%	.0%	1.0%	1.1%	

Note. Bivariate *r* coefficients are presented above the diagonal. Common variance (squared *r*) percentages are presented below the diagonal.

Figure 1
Scattergram of ANCOVA Data

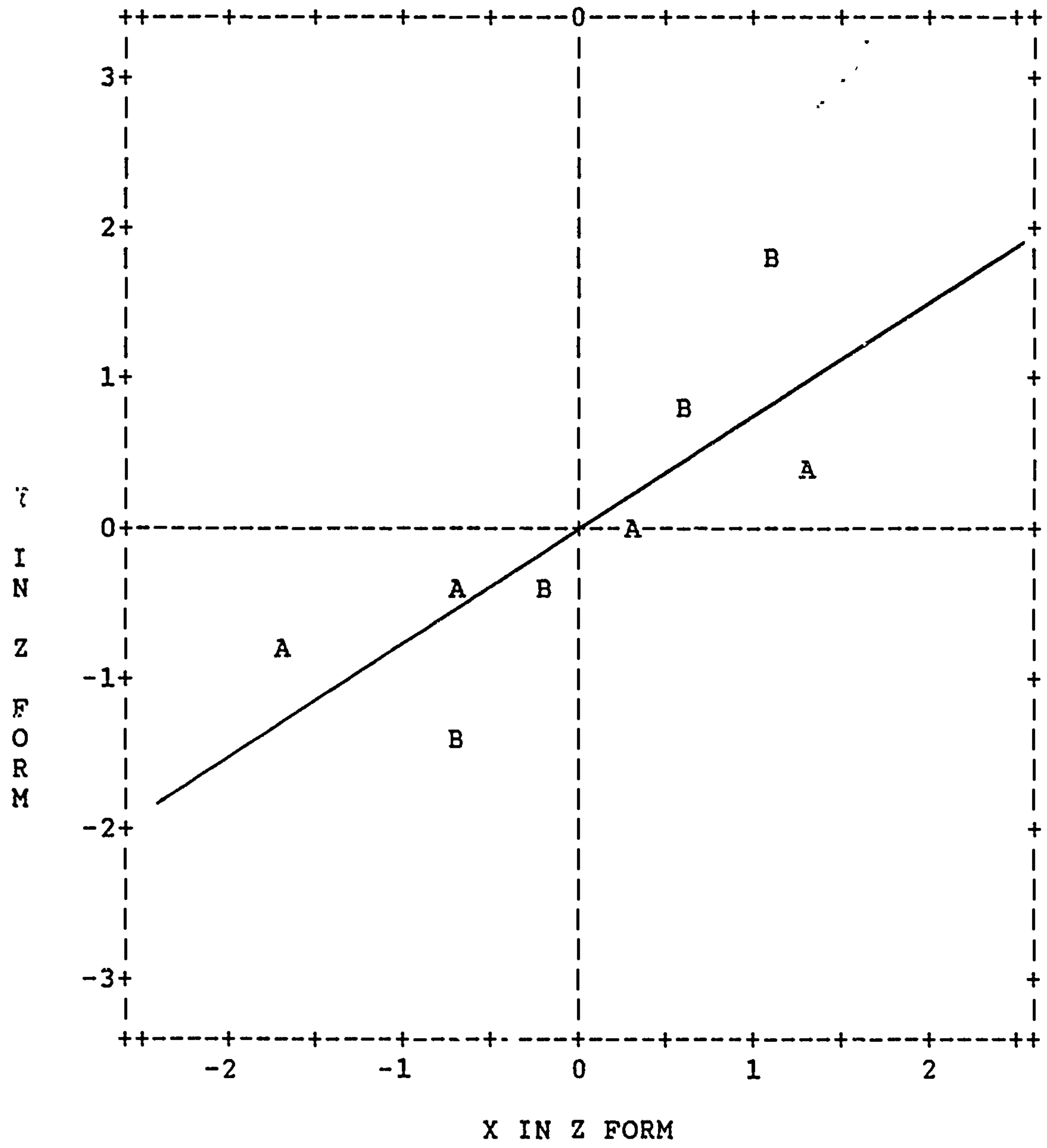

Figure 2

ANCOVA Best Case

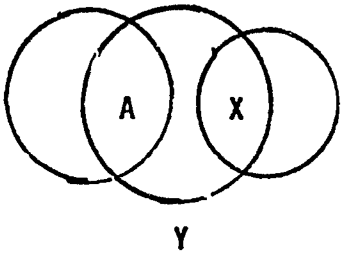

Figure 3

ANCOVA Worst Case