DOCUMENT RESUME

ED 429 110 TM 029 652

AUTHOR Thompson, Bruce
TITLE Common Methodology Mistakes in Educational Research, Revisited, along with a Primer on Both Effect Sizes and the Bootstrap.
PUB DATE 1999-04-00
NOTE 126p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Educational Research; *Effect Size; *Research Methodology; Research Problems; *Statistical Significance; Tables (Data)
IDENTIFIERS *Bootstrap Methods

ABSTRACT

As an extension of B. Thompson's 1998 invited address to the American Educational Research Association, this paper cites two additional common faux pas in research methodology and explores some research issues for the future. These two errors in methodology are the use of univariate analyses in the presence of multiple outcome variables (with the converse use of univariate analyses in post hoc explorations of detected multivariate effects) and the conversion of intervally scaled predictor variables into nominally scaled data in the service of the "of variance" (OVA) analyses. Among the research issues to receive further attention in the future is the appropriate use of statistical significance tests. The use of the descriptive bootstrap and the various types of effect size from which the researcher should select when characterizing quantitative results are also discussed. The paper concludes with an exploration of the conditions necessary and sufficient for the realization of improved practices in educational research. Three appendixes contain the Statistical Package for the Social Sciences for Windows syntax used to analyze data for three tables. (Contains 16 tables, 15 figures, and 173 references.) (SLD)

Common Methodology Mistakes in Educational Research, Revisited, Along with a Primer on both Effect Sizes and the Bootstrap

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Invited address presented at the annual meeting of the American Educational Research Association (session #44.25), Montreal, April 22, 1999. Justin Levitov first introduced me to the bootstrap, for which I remain most grateful. I also appreciate the thoughtful comments of Cliff Lunneborg and Russell Thompson on a previous draft of this paper. The author and related reprints may be accessed through Internet URL: "**http://acs.tamu.edu/~bbt6147/**".

Abstract

The present AERA invited address was solicited to address the theme for the 1999 annual meeting, "On the Threshold of the Millennium: Challenges and Opportunities." The paper represents an extension of my 1998 invited address, and cites two additional common methodology faux pas to complement those enumerated in the previous address. The remainder of these remarks are forward-looking. The paper then considers (a) the proper role of statistical significance tests in contemporary behavioral research, (b) the utility of the descriptive bootstrap, especially as regards the use of "modern" statistics, and (c) the various types of effect sizes from which researchers should be expected to select in characterizing quantitative results. The paper concludes with an exploration of the conditions necessary and sufficient for the realization of improved practices in educational research.

Introduction

In 1993, Carl Kaestle, prior to his term as President of the National Academy of Education, published in the Educational Researcher an article titled, "The Awful Reputation of Education Research." It is noteworthy that the article took as a given the conclusion that educational research suffers an awful reputation, and rather than justifying this conclusion, Kaestle focused instead on exploring the etiology of this reality. For example, Kaestle (1993) noted that the education R&D community is seemingly in perpetual disarray, and that there is a

> ...lack of consensus--lack of consensus on goals, lack of consensus on research results, and lack of a united front on funding priorities and procedures.... [T]he lack of consensus on goals is more than political; it is the result of a weak field that cannot make tough decisions to do some things and not others, so it does a little of everything... (p. 29)

Although Kaestle (1993) did not find it necessary to provide a warrant for his conclusion that educational research has an awful reputation, others have directly addressed this concern.

The National Academy of Science evaluated educational research generically, and found "methodologically weak research, trivial studies, an infatuation with jargon, and a tendency toward fads with a consequent fragmentation of effort" (Atkinson & Jackson, 1992, p. 20). Others also have argued that "too much of what we see in print is seriously flawed" as regards research methods, and that "much of the work in print ought not to be there" (Tuckman, 1990,

p. 22). Gall, Borg and Gall (1996) concurred, noting that "the quality of published studies in education and related disciplines is, unfortunately, not high" (p. 151).

Indeed, empirical studies of published research involving methodology experts as judges corroborate these impressions. For example, Hall, Ward and Comer (1988) and Ward, Hall and Schramm (1975) found that over 40% and over 60%, respectively, of published research was seriously or completely flawed. Wandt (1967) and Vockell and Asher (1974) reported similar results from their empirical studies of the quality of published research. Dissertations, too, have been examined, and have been found methodologically wanting (cf. Thompson, 1988a, 1994a).

Researchers have also questioned the ecological validity of both quantitative and qualitative educational studies. For example, Elliot Eisner studied two volumes of the flagship journal of the American Educational Research Association, the American Educational Research Journal (*AERJ*). He reported that,

> The median experimental treatment time for seven of the 15 experimental studies that reported experimental treatment time in Volume 18 of the *AERJ* is 1 hour and 15 minutes. I suppose that we should take some comfort in the fact that this represents a 66 percent increase over a 3-year period. In 1978 the median experimental treatment time per subject was 45 minutes. (Eisner, 1983, p. 14)

Similarly, Fetterman (1982) studied major qualitative projects, and reported that, "In one study, labeled 'An ethnographic study of..., observers were on site at only one point in time for five days. In a[nother] national study purporting to be ethnographic, once-a-week, on-site observations were made for 4 months" (p. 17)

None of this is to deny that educational research, whatever its methodological and other limits, has influenced and informed educational practice (cf. Gage, 1985; Travers, 1983). Even a methodologically flawed study may still contribute something to our understanding of educational phenomena. As Glass (1979) noted, "Our research literature in education is not of the highest quality, but I suspect that it is good enough on most topics" (p. 12).

However, as I pointed out in a 1998 AERA invited address, the problem with methodologically flawed educational studies is that these flaws are entirely *gratuitous*. I argued that

> incorrect analyses arise from doctoral methodology *instruction* that teaches research methods as series of rotely-followed routines, as against thoughtful elements of a reflective enterprise; from doctoral *curricula* that seemingly have less and less room for quantitative statistics and measurement content, even while our knowledge base in these areas is burgeoning (Aiken, West, Sechrest, Reno, with Roediger, Scarr, Kazdin & Sherman, 1990; Pedhazur & Schmelkin, 1991, pp. 2-3); and, in some cases, from an unfortunate *atavistic impulse* to somehow escape responsibility for analytic decisions by justifying choices, sans rationale, solely on the basis that

> the choices are common or traditional. (Thompson, 1998a, p. 4)

Such concerns have certainly been voiced by others. For example, following the 1998 annual AERA meeting, one conference attendee wrote AERA President Alan Schoenfeld to complain that

> At [the 1998 annual meeting] we had a hard time finding rigorous research that reported actual conclusions. Perhaps we should rename the association the American Educational Discussion Association.... This is a serious problem. By encouraging anything that passes for inquiry to be a valid way of discovering answers to complex questions, we support a culture of intuition and artistry rather than building reliable research bases and robust theories. Incidentally, theory was even harder to find than good research. (Anonymous, 1998, p. 41)

Subsequently, Schoenfeld appointed a new AERA committee, the Research Advisory Committee, which currently is chaired by Edmund Gordon. The current members of the Committee are: Ann Brown, Gary Fenstermacher, Eugene Garcia, Robert Glaser, James Greeno, Margaret LeCompte, Richard Shavelson, Vanessa Siddle Walker, and Alan Schoenfeld, ex officio, Lorrie Shepard, ex officio, and William Russell, ex officio. The Committee is charged to strengthen the research-related capacity of AERA and its members, coordinate its activities with appropriate AERA programs, and be entrepreneurial in nature. [In some respects, the AERA Research Advisory Committee has a mission similar to that of the APA Task Force on Statistical Inference, which was appointed in 1996 (Azar, 1997; Shea, 1996).]

AERA President Alan Schoenfeld also appointed Geoffrey Saxe the 1999 annual meeting program chair. Together, they then described the theme for the AERA annual meeting in Montreal:

> As we thought about possible themes for the upcoming annual meeting, we were pressed by a sense of timeliness and urgency. With regard to timeliness, ...the calendar year for the next annual meeting is 1999, the year that heralds the new millennium.... It's a propitious time to think about what we know, what we need to know, and where we should be heading. Thus, our overarching theme [for the 1999 annual meeting] is "On the Threshold of the Millennium: Challenges and Opportunities."
>
> There is also a sense of urgency. Like many others, we see the field of education at a point of critical choices--in some arenas, one might say crises. (Saxe & Schoenfeld, 1998, p. 41)

The present paper was among those invited by various divisions to address this theme, and is an extension of my 1998 AERA address (Thompson, 1998a).

Purpose of the Present Paper

In my 1998 AERA invited address I advocated the improvement of educational research via the eradication of five identified *faux pas*:

(1) the use of *stepwise* methods;

(2) the failure to consider in result interpretation the *context specificity* of analytic weights (e.g., regression beta weights, factor pattern coefficients, discriminant function coefficients, canonical function coefficients) that are part of all parametric quantitative analyses;

(3) the failure to interpret both *weights and structure coefficients* as part of result interpretation;

(4) the failure to recognize that *reliability* is a characteristic of scores, and not of tests; and

(5) the incorrect interpretation of *statistical significance* and the related failure to report and interpret the *effect sizes* present in all quantitative analyses.

Two Additional Methodology *Faux Pas*

The present didactic essay elaborates two additional common methodology errors to delineate a constellation of seven cardinal sins of analytic research practice:

(6) the use of univariate analyses in the presence of multiple outcomes variables, and the converse use of univariate analyses in post hoc explorations of detected *multivariate effects*; and

(7) the *conversion of intervally-scaled predictor variables* into nominally-scaled data in service of OVA (i.e., ANOVA, ANCOVA, MANOVA, MANCOVA) analyses.

However, the present paper is more than a further elaboration of bad behaviors. Here the discussion of these two errors focuses on driving home two important realizations that should undergird best methodological practice:

1. All statistical analyses of scores on measured/observed variables actually focus on correlational analyses of scores on synthetic/latent variables derived by applying weights to the observed variables; and
2. The researcher's fundamental task in deriving defensible results is to employ an analytic model that matches the researcher's (too often implicit) model of reality.

These two realization will provide a <u>conceptual foundation</u> for the treatment in the remainder of the paper.

<u>Focus on the Future: Improving Educational Research</u>

Although the focus on common methodological *faux pas* has some merit, in keeping with the theme of this 1999 annual meeting of AERA, the present invited address then turns toward the constructive portrayal of a brighter research future. Three issues are addressed. First, the proper role of *statistical significance* testing in future practice is explored. Second, the use of so-called "internal replicability" analyses in the form of the *bootstrap* is described. As part of this discussion some "modern" statistics are briefly discussed. Third, the computation and interpretation of *effects sizes* are described.

Other methods *faux pas* and other methods improvements might both have been elaborated. However, the proposed changes would result in considerable improvement in future educational research. In my view, (a) informed use of statistical tests, (b) the more frequent use of external and internal replicability analyses, and especially (c) required reporting and interpretation of effect sizes in all quantitative research are both necessary and sufficient conditions for realizing improvements.

Essentials for Realizing Improvements

The essay ends by considering how fields move and what must be done to realize these potential improvements. In my view, AERA must exercise visible and coherent academic leadership if change is to occur. To date, such leadership has not often been within the organization's traditions.

Faux Pas #6: Univariate as Against Multivariate Analyses

Too often, educational researchers invoke a series of univariate analyses (e.g., ANOVA, regression) to analyze multiple dependent variable scores from a single sample of participants. Conversely, too often researchers who correctly select a multivariate analysis invoke univariate analyses *post hoc* in their investigation of the origins of multivariate effects. Here it will be demonstrated once again, using heuristic data to make the discussion completely concrete, that in both cases these choices may lead to serious interpretation errors.

The fundamental conceptual emphasis of this discussion, as previously noted, is on making the point that:

1. *All statistical analyses of scores on measured/observed variables actually focus on correlational analyses of scores on synthetic/latent variables derived by applying weights to the observed variables.*

Two small heuristic data sets are employed to illustrate the relevant dynamics, respectively, for the univariate (i.e., single dependent/outcome variable) and multivariate (i.e., multiple outcome variables) cases.

Univariate Case

Table 1 presents a heuristic data set involving scores on three measured/observed variables: Y, X1, and X2. These variables are called "measured" (or "observed") because they are directly measured, *without any application of additive or multiplicative weights*, via rulers, scales, or psychometric tools.

INSERT TABLE 1 ABOUT HERE.

However, ALL parametric analyses apply weights to the measured/observed variables to estimate scores for each person on synthetic or latent variables. This is true notwithstanding the fact that for some statistical analyses (e.g., ANOVA) the weights are not printed by some statistical packages. As I have noted elsewhere, the weights in different analyses

> ...are all analogous, but are given different names in different analyses (e.g., beta weights in regression, pattern coefficients in factor analysis, discriminant function coefficients in discriminant analysis, and canonical function coefficients in canonical correlation analysis), mainly to obfuscate the commonalities of [all] parametric methods, and to confuse graduate students. (Thompson, 1992a, pp. 906-907)

The synthetic variables derived by applying weights to the measured variables then become the focus of the statistical analyses.

The fact that all analyses are part on one single General Linear Model (GLM) family is a fundamental foundational understanding essential (in my view) to the informed selection of analytic methods. The seminal readings have been provided by Cohen (1968) viz. the univariate case, by Knapp (1978) viz. the multivariate case, and by Bagozzi, Fornell and Larcker (1981) regarding the most general case of the GLM: structural equation modeling. Related heuristic demonstrations of General Linear Model dynamics have been offered by Fan (1996, 1997) and Thompson (1984, 1991, 1998a, in press-a).

In the multiple regression case, a given i_{th} person's score on the measured/observed variable $\underline{Y}_i$ is estimated as the synthetic/latent variable $\hat{\underline{Y}}_i$. The predicted outcome score for a given person equals $\hat{\underline{Y}}_i = a + b_1(\underline{X1}_i) + b_2(\underline{X2}_i)$, which for these data, as reported in Figure 1, equals $-581.735382 + [1.301899 \times X1_i] + [0.862072 \times X2_i]$. For example, for person 1, $\hat{\underline{Y}}_1 = [1.301899 \times 392] + [0.862072 \times 573] = 422.58$.

INSERT FIGURE 1 ABOUT HERE.

Some Noteworthy Revelations. The "ordinary least squares" (OLS) estimation used in classical regression analysis optimizes the fit in the sample of each $\hat{\underline{Y}}_i$ to each $\underline{Y}_i$ score. Consequently, as noted by Thompson (1992b), even if all the predictors are useless, the means of $\hat{\underline{Y}}$ and $\underline{Y}$ will always be equal (here 500.25), and the mean of the $\underline{e}$ scores ($\underline{e}_i = \underline{Y}_i - \hat{\underline{Y}}_i$) will always be zero. These expectations are confirmed in the Table 1 results.

It is also worth noting that the sum of squares (i.e., the sum of the squared deviations of each person's score from the mean) of the $\hat{\underline{Y}}$ scores (i.e., 167,218.50) computed in Table 1 matches the "regression" sum of squares (variously synonymously called "explained," "model," "between," so as to confuse the graduate students) reported in the Figure 1 SPSS output. Furthermore, the sum of squares of the e scores reported in Table 1 (i.e., 32,821.26) exactly matches the "residual" sum of squares (variously called "error," "unexplained," and "residual") value reported in the Figure 1 SPSS output.

It is especially noteworthy that the sum of squares explained (i.e., 167,218.50) divided the sum of squares of the Y scores (i.e., the sum of squares "total" = 167,218.50 + 32,821.26 = 200,039.75) tells us the proportion of the variance in the Y scores that we can predict given knowledge of the X1 and the X2 scores. For these data the proportion is 167,218.50 / 200,039.75 = .83593. This formula is one of several formulas with which to compute the uncorrected regression effect size, the multiple R^2.

Indeed, for the univariate case, because ALL analyses are correlational, an r^2 analog of this effect size can always be computed, using this formula across analyses. However, in ANOVA, for example, when we compute this effect size using this generic formula, we call the result eta^2 (η^2; or synonymously the correlation ratio [not the correlation coefficient!]), primarily to confuse the graduate students.

Even More Important Revelations. Figure 2 presents the correlation coefficients involving all possible pairs of the five (three measured, two synthetic) variables. Several additional revelations become obvious.

INSERT FIGURE 2 ABOUT HERE.

First, note that the $\hat{\underline{Y}}$ scores and the $\underline{e}$ scores are perfectly uncorrelated. This will ALWAYS be the case, by definition, since the $\hat{\underline{Y}}$ scores are the aspects of the $\underline{Y}$ scores that the predictors can explain or predict, and the $\underline{e}$ scores are the aspects of the $\underline{Y}$ scores that the predictors cannot explain or predict (i.e., because $\underline{e}_i$ is defined as $\underline{Y}_i - \hat{\underline{Y}}_i$, therefore $\underline{r}_{\text{YHAT x e}} = 0$). Similarly, the measured predictor variables (here $\underline{X1}$ and $\underline{X2}$) always have correlations of zero with the $\underline{e}$ scores, again because the $\underline{e}$ scores by definition are the parts of the $\underline{Y}$ scores that the predictors cannot explain.

Second, note that the $\underline{r}_{\text{Y x YHAT}}$ reported in Figure 3 (i.e., .9143) matches the multiple $\underline{R}$ reported in Figure 1 (i.e., .91429), except for the arbitrary decision by different computer programs to present these statistics to different numbers of decimal places. The equality makes sense conceptually, if we think of the $\hat{\underline{Y}}$ scores as being the part of the predictors useful in predicting/explaining the $\underline{Y}$ scores, discarding all the parts of the measured predictors that are not useful (about which we are completely uninterested, because the focus of the analysis is solely on the outcome variable).

This last revelation is <u>extremely</u> important to a conceptual understanding of statistical analyses. The fact that $\underline{R}_{\text{Y with X1, X2}} = \underline{r}_{\text{Y x YHAT}}$ means that the synthetic variable, $\hat{\underline{Y}}$, is actually the focus of the analysis. Indeed, synthetic variables are ALWAYS the real focus of statistical analyses!

This makes sense, when we realize that our measures are only indicators of our psychological constructs, and that what we really care about in educational research are not the observed scores on our measurement tools *per se*, but instead is the underlying construct. For example, if I wish to improve the self-concepts of third-grade elementary students, what I really care about is improving their unobservable self-concepts, and not the scores on an imperfect measure of this construct, which I only use as a vehicle to estimate the latent construct of interest, because the construct cannot be directly observed.

Third, the correlations of the measured predictor variables with the synthetic variable (i.e., .7512 and -.0741) are called "structure" coefficients. These can also be derived by computation (cf. Thompson & Borrello, 1985) as $\underline{r}_S = \underline{r}_{Y \text{ with } X}$ / $\underline{R}$ (e.g., .6868 / .91429 = .7512). [Due to a strategic error on the part of methodology professors, who convene annually in a secret coven to generate more statistical terminology with which to confuse the graduate students, for some reason the mathematically analogous structure coefficients across all analyses are uniformly called by the same name--an oversight that will doubtless soon be corrected.]

The reason structure coefficients are called "structure" coefficients is that these coefficients provide insight regarding what is the nature or the structure of the underlying synthetic variables of the actual research focus. Although space precludes further detail here, I regard the interpretation of structure coefficients are being essential in most research applications (Thompson, 1997b, 1998a; Thompson & Borrello, 1985). Some educational researchers erroneously believe that these coefficients are unimportant insofar as they are not reported for all analyses by some computer packages; these researchers incorrectly believe that SPSS and other computer packages were written in a sole authorship venture by a benevolent God who has elected judiciously to report on printouts (a) *all* results of interest and (b) *only* the results of genuine interest.

The Critical, Essential Revelation. Figure 2 also provides the basis for delineating a paradox which, once resolved, leads to a fundamentally important insight regarding statistical analyses. Notice for these data the $\underline{r}^2$ between $\underline{Y}$ and $\underline{X1}$ is $.6868^2 = 47.17\%$ and the $\underline{r}^2$ between $\underline{Y}$ and $\underline{X2}$ is $-.0677^2 = 0.46\%$. The sum of these two values is .4763.

Yet, as reported in Figures 2 and 3, the $\underline{R}^2$ value for these data is $.91429^2 = 83.593\%$, a value approaching the mathematical limen for $\underline{R}^2$. How can the multiple $\underline{R}^2$ value (83.593%) be not only larger, but nearly twice as large as the sum of the $\underline{r}^2$ values of the two predictor variables with $\underline{Y}$?

These data illustrate a "suppressor" effect. These effects were first noted in World War II when psychologists used paper-and-pencil measures of spatial and mechanical ability to predict ability to pilot planes. Counterintuitively, it was discovered that verbal ability, which is essentially unrelated with pilot ability, nevertheless substantially improved the $\underline{R}^2$ when used as a predictor in conjunction spatial and mechanical ability scores. As Horst (1966, p. 355) explained, "To include the verbal score with a negative weight served to suppress or subtract irrelevant [measurement artifact] ability [in the spatial and mechanical ability scores], and to discount the scores of those who did well on the test simply because of their verbal ability rather than because of abilities required for success in pilot training."

Thus, suppressor effects are desirable, notwithstanding what some may deem a pejorative name, because suppressor effects actually increase effect sizes. Henard (1998) and Lancaster (in press) provide readable elaborations. All this discussion leads to the extremely important point that

> **The latent or synthetic variables analyzed in all parametric methods are always more than the sum of their constituent parts**. If we only look at observed variables, such as by only examining a series of bivariate _r_'s, we can easily under or overestimate the actual effects that are embedded within our data. We must use analytic methods that honor the complexities of the reality that we purportedly wish to study--a reality in which variables can interact in all sorts of complex and counterintuitive ways. (Thompson, 1992b, pp. 13-14, emphasis in original)

Multivariate Case

Table 2 presents heuristic data for 10 people in each of two groups on two measured/observed outcome/response variables, _X_ and _Y_. These data are somewhat similar to those reported by Fish (1988), who argued that multivariate analyses are usually vital. The Table 2 data are used here to illustrate that (a) when you have more than one outcome variable, multivariate analyses may be essential, and (b) when you do a multivariate analysis, you must not use a univariate method *post hoc* to explore the detected multivariate effects.

INSERT TABLE 2 ABOUT HERE.

For these heuristic data, the outcome scores of X and Y have exactly the same variance in both groups 1 and 2, as reported in the bottom of Table 2. This exactly equal SD (and variance and sum of squares) means that the ANOVA "homogeneity of variance" assumption (called this because this characterization sounds fancier than simply saying "the outcome variable scores were equally 'spread out' in all groups") was perfectly met, and therefore the calculated ANOVA F test results are exactly accurate for these data. Furthermore, the analogous multivariate "homogeneity of dispersion matrices" assumption (meaning simply that the variance/covariance matrices in the two groups were equal) was also perfectly met, and therefore the MANOVA F tests are exactly accurate as well. In short, the demonstrations here are not contaminated by the failure to meet statistical assumptions!

Figure 3 presents ANOVA results for separate analyses of the X and Y scores presented in Table 2. For both X and Y, the two means do not differ to a statistically significant degree. In fact, for both variables the $p_{\text{CALCULATED}}$ values were .774. Furthermore, the eta^2 effect sizes were both computed to be 0.469% (e.g., 5.0 / [5.0 + 1061.0] = 5.0 / 1065.0 = .00469). Thus, the two sets of ANOVA results are not statistically significant and they both involve extremely small effect sizes.

INSERT FIGURE 3 ABOUT HERE.

However, as also reported in the Figure 3 results, a MANOVA/Descriptive Discriminant Analysis (DDA; for a one-way MANOVA, MANOVA and DDA yield the same results, but the DDA provides more detailed analysis--see Huberty, 1994; Huberty & Barton, 1989; Thompson, 1995b) of the *same data* yields a $p_{\text{CALCULATED}}$ value of .000239, and an eta^2 of 62.5%. Clearly, the resulting interpretation of the same data would be night-and-day different for these two sets of analyses. Again, the synthetic variables in some senses can become more than the sum of their parts, as was also the case in the previous heuristic demonstration.

Table 2 reports these latent variable scores for the 20 participants, derived by applying the weights (-1.225 and 1.225) reported in Figure 3 to the two measured outcome variables. For heuristic purposes only, the scores on the synthetic variable labelled "DSCORE" were then subjected to the ANOVA reported in Figure 4. As reported in Figure 4, this analysis of the multivariate synthetic variable, a weighted aggregation of the outcome variables $\underline{X}$ and $\underline{Y}$, yields the same eta^2 effect size (i.e., 62.5%) reported in Figure 3 for the DDA/MANOVA results. Again, all statistical analyses actually focus on the synthetic/latent variables actually derived in the analyses, *quod erat demonstrandum*.

INSERT FIGURE 4 ABOUT HERE.

The present heuristic example can be framed in either of two ways, both of which highlight common errors in contemporary analytic practice. The first error involves conducting multiple univariate analyses to evaluate multivariate data; the second error involves using univariate analyses (e.g., ANOVAs) in *post hoc* analyses of detected multivariate effects.

Using Several Univariate Analyses to Analyze Multivariate Data. The present example might be framed as an illustration of a researcher conducting *only* two ANOVAs to analyze the two sets of dependent variable scores. The researcher here would find no statistically significant (both $\underline{p}_{CALCULATED}$ values = .774) nor (probably, depending upon the context of the study and researcher personal values) any noteworthy effect (both eta^2 values = 0.469%). This researcher would remain oblivious to the statistically significant effect ($\underline{p}_{CALCULATED}$ = .000239) and huge (as regards typicality; see Cohen, 1988) effect size (multivariate eta^2 = 62.5%).

One potentially noteworthy argument in favor of employing multivariate methods with data involving more than one outcome variable involves the inflation of "experimentwise" Type I error rates (α_{EW}; i.e., the probability of making one or more Type I errors in a set of hypothesis tests--see Thompson, 1994d). At the extreme, when the outcome variables or the hypotheses (as in a balanced ANOVA design) are perfectly uncorrelated, α_{EW} is a function of the "testwise" alpha level (α_{TW}) and the number of outcome variables or hypotheses tested ($\underline{k}$), and equals

$$1 - (1 - \alpha_{TW})^k.$$

Because this function is exponential, experimentwise error rates can inflate quite rapidly! [Imagine my consternation when I detected a local dissertation invoking more than 1,000 univariate statistical significance tests (Thompson, 1994a).]

One way to control the inflation of experimentwise error is to use a "Bonferroni correction" which adjusts the α_{TW} downward so as to minimize the final α_{EW}. Of course, one consequence of this strategy is lessened statistical power against Type II error. However, the primary argument against using a series of univariate analyses to evaluate data involving multiple outcome variables does not invoke statistical significance testing concepts.

Multivariate methods are often vital in behavioral research simply because *multivariate methods best honor the reality to which the researcher is purportedly trying to generalize*. Implicit within every analysis is an analytic model. Each researcher also has a presumptive model of what reality is believed to be like. It is critical that our analytic models and our models of reality match, otherwise our conclusions will be invalid. It is generally best to consciously reflect on the fit of these two models whenever we do research. Of course, researchers with different models of reality may make different analytic choices, but this is not disturbing because analytic choices are philosophically driven anyway (Cliff, 1987, p. 349).

My personal model of reality is one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects" (Thompson, 1986b, p. 9). Given such a model of reality, it is critical that the full network of all possible relationships be considered *simultaneously* within the analysis. Otherwise, the Figure 3 multivariate effects, presumptively real given my model of reality, would go undetected. Thus, Tatsuoka's (1973b) previous remarks remain telling:

> The often-heard argument, "I'm more interested in seeing how each variable, in its own right, affects the outcome" overlooks the fact that any variable taken in isolation may affect the criterion differently from the way it will act in the company of other variables. It also overlooks the fact that multivariate analysis--precisely by considering all the variables simultaneously--can throw light on how each one contributes to the relation. (p. 273)

For these various reasons empirical studies (Emmons, Stallings & Layne, 1990) show that, "In the last 20 years, the use of multivariate statistics has become commonplace" (Grimm & Yarnold, 1995, p. vii).

Using Univariate Analyses *post hoc* to Investigate Detected Multivariate Effects. In ANOVA and ANCOVA, *post hoc* (also called "*a posteriori*," "unplanned," and "unfocused") contrasts (also called "comparisons") are necessary to explore the origins of detected omnibus effects iff ("if and only if") (a) an omnibus effect is statistically significant (but see Barnette & McLean, 1998) and (b) the way (also called an OVA "factor", but this alternative name tends to become confused with a factor analysis "factor") has more than two levels.

However, in MANOVA and MANCOVA *post hoc* tests are necessary to evaluate (a) which groups differ (b) as regards which one or more outcome variables. Even in a two-level way (or "factor"), if the effect is statistically significant, further analyses are necessary to determine on which one or more outcome/response variables the two groups differ. An alarming number of researchers employ ANOVA as a *post hoc* analysis to explore detected MANOVA effects (Thompson, 1999b).

Unfortunately, as the previous example made clear, because the two *post hoc* ANOVAs would fail to explain where the incredibly large and statistically significant MANOVA effect originated, ANOVA is not a suitable MANOVA *post hoc* analysis. As Borgen and Seling (1978) argued, "When data truly are multivariate, as implied by the application of MANOVA, a multivariate follow-up technique seems necessary to 'discover' the complexity of the data" (p. 696). It is simply illogical to first declare interest in a multivariate omnibus system of variables, and to then explore detected effects in this multivariate world by conducting non-multivariate tests!

Faux Pas #7: Discarding Variance in Intervally-Scaled Variables

Historically, OVA methods (i.e., ANOVA, ANCOVA, MANOVA, MANCOVA) dominated the social scientist's analytic landscape (Edgington, 1964, 1974). However, more recently the proportion of uses of OVA methods has declined (cf. Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980). Planned contrasts (Thompson, 1985, 1986a, 1994c) have been increasingly favored over omnibus tests. And regression and related techniques within the GLM family have been increasingly employed.

Improved analytic choices have partially been a function of growing researcher awareness that:

2. *The researcher's fundamental task in deriving defensible results is to employ an analytic model that matches the researcher's (too often implicit) model of reality.*

This growing awareness can largely be traced to a seminal article written by Jacob Cohen (1968, p. 426).

Theory

Cohen (1968) noted that ANOVA and ANCOVA are special cases of multiple regression analysis, and argued that in this realization "lie possibilities for more relevant and therefore more powerful exploitation of research data." Since that time researchers have increasingly recognized that conventional multiple regression analysis of data as they were initially collected (no conversion of intervally scaled independent variables into dichotomies or trichotomies) does not discard information or distort reality, and that the "general linear model"

> ...can be used equally well in experimental or non-experimental research. It can handle continuous and categorical variables. It can handle two, three, four, or more independent variables... Finally, as we will abundantly show, multiple regression analysis can do anything the analysis of variance does--sums of squares, mean squares, F ratios--and more. (Kerlinger & Pedhazur, 1973, p. 3)

Discarding variance is generally not good research practice. As Kerlinger (1986) explained,

> ...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. A good rule of research data analysis, therefore, is: <u>Do not reduce continuous variables to partitioned variables</u> (dichotomies, trichotomies, etc.) unless compelled to do so by circumstances or the nature of the data (seriously skewed, bimodal, etc.). (p. 558, emphasis in original)

Kerlinger (1986, p. 558) noted that variance is the "stuff" on which all analysis is based. Discarding variance by categorizing intervally-scaled variables amounts to the "squandering of information" (Cohen, 1968, p. 441). As Pedhazur (1982, pp. 452-453) emphasized,

> Categorization of attribute variables is all too frequently resorted to in the social sciences.... It is possible that some of the conflicting evidence in the research literature of a given area may be attributed to the practice of categorization of continuous variables.... Categorization leads to a loss of information, and consequently to a less sensitive analysis.

Some researchers may be prone to categorizing continuous variables and overuse of ANOVA because they <u>unconsciously</u> and <u>erroneously</u> associate ANOVA with the power of experimental designs. As I have noted previously,

> Even most experimental studies invoke intervally scaled "aptitude" variables (e.g., IQ scores in a study with academic achievement as a dependent variable), to conduct the aptitude-treatment interaction (ATI) analyses recommended so persuasively by Cronbach (1957, 1975) in his 1957 APA Presidential address. (Thompson, 1993a, pp. 7-8)

Thus, many researchers employ interval predictor variables, even in experimental designs, but these same researchers too often convert their interval predictor variables to nominal scale merely to conduct OVA analyses.

It is *true* that experimental designs allow causal inferences and that ANOVA is appropriate for many experimental designs. However, it is *not* therefore *true* that doing an ANOVA makes the design experimental and thus allows causal inferences.

Humphreys (1978, p. 873, emphasis added) noted that:

> The basic fact is that a measure of individual differences is not an independent variable [in a experimental design], and it *does not become one* by categorizing the scores and treating the categories as if they defined a variable under experimental control in a factorially designed analysis of variance.

Similarly, Humphreys and Fleishman (1974, p. 468) noted that categorizing variables in a nonexperimental design using an ANOVA analysis "not infrequently produces in both the investigator and his audience the illusion that he has experimental control over the independent variable. Nothing could be more wrong." Because within the general linear model all analyses are correlational, and it is the design and not the analysis that yields the capacity to make causal inferences, the practice of converting intervally-scaled predictor variables to nominal scale so that ANOVA and other OVAs (i.e., ANCOVA, MANOVA, MANCOVA) can be conducted is inexcusable, at least in most cases.

As Cliff (1987, p. 130, emphasis added) noted, the practice of discarding variance on intervally-scaled predictor variables to perform OVA analyses creates problems in almost all cases:

> 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 [or a trichotomy] makes the variable more unreliable, not less.

In such cases, it is the reliability of the dichotomy that we actually analyze, and not the reliability of the highly-reliable, intervally-scaled data that we originally collected, which impact the analysis we are actually conducting.

Heuristic Examples for Three Possible Cases

When we convert an intervally-scaled independent variable into a nominally-scaled way in service of performing an OVA analysis, we are implicitly invoking a model of reality with two strict assumptions:

1. all the participants assigned to a given level of the way (or "factor") are the same, and
2. all the participants assigned to different levels of the way are different.

For example, if we have a normal distribution of IQ scores, and we use scores of 90 and 110 to trichotomize our interval data, we are saying that:

1. the 2 people in the High IQ group with IQs of 111 and 145 are the same, and
2. the 2 people in the Low and Middle IQ groups with IQs of 89 and 91, respectively, are different.

Whether our decision to convert our intervally-scaled data to nominal scale is appropriate depends entirely on the research situation. There are three possible situations.

Table 3 presents heuristic data illustrating the three possibilities. The measured/observed outcome variable in all three cases is _Y_.

INSERT TABLE 3 ABOUT HERE.

Case #1: No harm, no foul. In case #1 the intervally-scaled variable _X1_ is re-expressed as a trichotomy in the form of variable _X1'_. Assuming that the standard error of the measurement is something like 3 or 6, the conversion in this instance does not seem problematic, because it appears reasonable to assume that:

1. all the participants assigned to a given level of the way are the same, and
2. all the participants assigned to different levels of the way

are different.

Case #2: Creating variance where there is none. Case #2 again assumes that the standard error of the measurement is something like 3 to 6 for the hypothetical scores. Here none of the 21 participants appear to be different as regards their scores on Table 3 variable X2, so assigning the participants to three groups via variable X2' seems to create differences where there are none. This will generate analytic results in which the analytic model does not honor our model of reality, which in turn compromises the integrity of our results.

Some may protest that no real researcher would ever, ever assign people to groups where there are, in fact, no meaningful differences among the participants as regards their scores on an independent variable. But consider a recent local dissertation that involved administration of a depression measure to children; based on scores on this measure the children were assigned to one of three depression groups. Regrettably, these children were all apparently happy and well-adjusted.

> It is especially interesting that the highest score on this [depression] variable... was apparently 3.43 (p. 57). As... [the student] acknowledged, the PNID authors themselves recommend a cutoff score of 4 for classifying subjects as being severely depressed. Thus, the highest score in... [the] entire sample appeared to be less than the minimum cutoff score suggested by the test's own authors! (Thompson, 1994a, p. 24)

Case #3: Discarding variance, distorting distribution shape. Alternatively, presume that the intervally-scaled independent variable (e.g., an aptitude way in an ATI design) is somewhat normally distributed. Variable X3 in Table 3 can be used to illustrate the potential consequences of re-expressing this information in the form of a nominally-scaled variable such as X3′.

Figure 5 presents the SPSS output from analyzing the data in both unmutilated (i.e., X3) and mutilated (i.e., X3′) form. In unmutilated form, the results are statistically significant ($p_{CALCULATED}$ = .00004) and the R^2 effect size is 59.7%. For the mutilated data, the results are not statistically significant at a conventional alpha level ($p_{CALCULATED}$ = .1145) and the eta^2 effect size is 21.4%, roughly a third of the effect for the regression analysis.

INSERT FIGURE 5 ABOUT HERE.

Criticisms of Statistical Significance Tests

Tenor of Past Criticism

The last several decades have delineated an exponential growth curve in the decade-by-decade criticisms across disciplines of statistical testing practices (Anderson, Burnham & Thompson, 1999). In their historical summary dating back to the origins of these tests, Huberty and Pike (in press) provide a thoughtful review of how we got to where we're at. Among the recent commentaries on statistical testing practices, I prefer Cohen (1994), Kirk (1996), Rosnow and Rosenthal (1989), Schmidt (1996), and Thompson (1996).

Among the classical criticisms, my favorites are Carver (1978), Meehl (1978), and Rozeboom (1960).

Among the more thoughtful works advocating statistical testing, I would cite Cortina and Dunlap (1997), Frick (1996), and especially Abelson (1997). The most balanced and comprehensive treatment is provided by Harlow, Mulaik and Steiger (1997) (for reviews of this book, see Levin, 1998 and Thompson, 1998c).

My purpose here is not to further articulate the various criticisms of statistical significance tests. My own recent thinking is elaborated in the several reports enumerated in Table 4. The focus here is on what should be the future. Therefore, criticisms of statistical tests are only briefly summarized in the present treatment.

INSERT TABLE 4 ABOUT HERE.

But two quotations may convey the tenor of some of these commentaries. Rozeboom (1997) recently argued that

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

And Tryon (1998) recently lamented,

> [T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses

> is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial... (p. 796)

Indeed, empirical studies confirm that many researchers do not fully understand the logic of their statistical tests (cf. Mittag, 1999; Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). Misconceptions are taught even in widely-used statistics textbooks (Carver, 1978).

Brief Summary of Four Criticisms of Common Practice

Statistical significance tests evaluate the probability of obtaining sample statistics (e.g., means, medians, correlation coefficients) that diverge as far from the null hypothesis as the sample statistics, or further, assuming that the null hypothesis is true in the population, and given the sample size (Cohen, 1994; Thompson, 1996). The utility of these estimates has been questioned on various grounds, four of which are briefly summarized here.

Conventionally, Statistical Tests Assume "Nil" Null Hypotheses. Cohen (1994) defined a "nil" null hypothesis as a null specifying no differences (e.g., H_0: $\underline{SD}_1 - \underline{SD}_2 = 0$) or zero correlations (e.g., $\underline{R}^2=0$). Researchers must specify some null hypothesis, or otherwise the probability of the sample statistics is completely indeterminate (Thompson, 1996)--infinitely many p values become equally plausible. But "nil" nulls are not required. Nevertheless, "as almost universally used, the null in H_0 is taken to mean nil, zero" (Cohen, 1994, p. 1000).

Some researchers employ nil nulls because statistical theory does not easily accommodate the testing of some non-nil nulls. But probably most researchers employ nil nulls because these nulls have been unconsciously accepted as traditional, because these nulls can be mindlessly formulated without consulting previous literature, or because most computer software defaults to tests of nil nulls (Thompson, 1998c, 1999a). As Boring (1919) argued 80 years ago, in his critique of the mindless use of statistical tests titled, "Mathematical vs. scientific significance,"

> The case is one of many where statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere. (p. 338)

I believe that when researchers presume a nil null is true in the population, an untruth is posited. As Meehl (1978, p. 822) noted, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the [nil] null hypothesis, taken literally, is always false." Similarly, Hays (1981, p. 293) pointed out that "[t]here is surely nothing on earth that is completely independent of anything else [in the population]. The strength of association may approach zero, but it should seldom or never be exactly zero." Roger Kirk (1996) concurred, noting that:

> It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers

> to focus on controlling the Type I error that cannot occur because *all* null hypotheses are false. (p. 747, emphasis added)

A $\underline{p}_{\text{CALCULATED}}$ value computed on the foundation of a false premise is inherently of somewhat limited utility. As I have noted previously, "in many contexts the use of a 'nil' hypothesis as the hypothesis we assume can render me largely disinterested in whether a result is 'nonchance'" (Thompson, 1997a, p. 30).

Particularly egregious is the use of "nil" nulls to test measurement hypotheses, where wildly non-nil results are both anticipated and demanded. As Abelson (1997) explained,

> And when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80'ish. (p. 121)

<u>Statistical Tests Can be a Tautological Evaluation of Sample Size</u>. When "nil" nulls are used, the null will always be rejected at some sample size. There are infinitely many possible sample effects. Given this, the probability of realizing an exactly zero sample effect is infinitely small. Therefore, given a "nil" null, and a non-zero sample effect, the null hypothesis will <u>always</u> be rejected at some sample size!

Consequently, as Hays (1981) emphasized, "virtually any study can be made to show significant results if one uses enough subjects" (p. 293). This means that

> Statistical significance testing can involve a

> tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. (Thompson, 1992c, p. 436)

Certainly this dynamic is well known, if it is just as widely ignored. More than 60 years ago, Berkson (1938) wrote an article titled, "Some difficulties of interpretation encountered in the application of the chi-square test." He noted that when working with data from roughly 200,000 people,

> an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the _P_'s tend to come out small... [W]e know in advance the _P_ that will result from an application of a chi-square test to a large sample... But since the result of the former is known, it is no test at all! (pp. 526-527)

Some 30 years ago, Bakan (1966) reported that, "The author had occasion to run a number of tests of significance on a battery of tests collected on about 60,000 subjects from all over the United States. Every test came out significant" (p. 425). Shortly thereafter, Kaiser (1976) reported not being surprised when many substantively trivial factors were found to be statistically significant when data were available from 40,000 participants.

Because Statistical Tests Assume Rather than Test the Population, Statistical Tests Do Not Evaluate Result Replicability. Too many researchers incorrectly assume, consciously or unconsciously, that the p values calculated in statistical significance tests evaluate the probability that results will replicate (Carver, 1978, 1993). But statistical tests do not evaluate the probability that the sample statistics occur in the population as parameters (Cohen, 1994).

Obviously, knowing the probability of the sample is less interesting than knowing the probability of the population. Knowing the probability of population parameters would bear upon result replicability, because we would then know something about the population from which future researchers would also draw their samples. But as Shaver (1993) argued so emphatically:

> [A] test of statistical significance is not an indication of the probability that a result would be obtained upon replication of the study.... Carver's (1978) treatment should have dealt a death blow to this fallacy.... (p. 304)

And so Cohen (1994) concluded that the statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997).

Statistical Significance Tests Do Not Solely Evaluate Effect Magnitude. Because various study features (including score reliability) impact calculated p values, $p_{\text{CALCULATED}}$ cannot be used as a satisfactory index of study effect size. As I have noted elsewhere,

> The calculated p values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because p values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (Thompson, 1999a, pp. 169-170)

The recent fourth edition of the American Psychological Association style manual (APA, 1994) explicitly acknowledged that p values are not acceptable indices of effect:

> Neither of the two types of probability values [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size... You are [therefore] *encouraged* to provide effect-size information. (APA, 1994, p. 18, emphasis added)

In short, effect sizes should be reported in every quantitative study.

The "Bootstrap"

Explanation of the "bootstrap" will provide a concrete basis for facilitating genuine understanding of what statistical tests do (and do not) do. The "bootstrap" has been so named because this statistical procedure represents an attempt to "pull oneself up" on one's own, using one's sample data, without external assistance from a theoretically-derived sampling distribution.

Related books have been offered by Davison and Hinkley (1997), Efron and Tibshirani (1993), Manly (1994), and Sprent (1998). Accessible shorter conceptual treatments have been presented by Diaconis and Efron (1983) and Thompson (1993b). I especially and particularly recommend the remarkable book by Lunneborg (1999).

Software to invoke the bootstrap is available in most structural equation modeling software (e.g., EQS, AMOS). Specialized bootstrap software for microcomputers (e.g., S Plus, SC, and Resampling Stats) is also readily available.

The Sampling Distribution

Key to understanding statistical significance tests is understanding the sample distribution and distinguishing the (a) sampling distribution from (b) the population distribution and (c) the score distribution. Among the better book treatments is one offered by Hinkle, Wiersma and Jurs (1998, pp. 176-178). Shorter treatments include those by Breunig (1995), Mittag (1992), and Rennie (1997).

The *population* distribution consists of the scores of the N entities (e.g., people, laboratory mice) of interest to the researcher, regarding whom the researcher wishes to generalize. In the social sciences, many researchers deem the population to be infinite. For example, an educational researcher may hope to generalize about the effects of a teaching method on all human beings across time.

Researchers typically describe the population by computing or estimating characterizations of the population scores (e.g., means, interquartile ranges), so that the population can be more readily comprehended. These characterizations of the population are called "parameters," and are conventionally symbolized using Greek letters (e.g., μ for the population score mean, σ for the population score standard deviation).

The *sample* distribution also consists of scores, but only a subsample of n scores from the population. The characterizations of the sample scores are called "statistics," and are conventionally represented by Roman letters (e.g., M, SD, r). Strictly speaking, statistical significance tests evaluate the probability of a given set of statistics occurring, assuming that the sample came from a population exactly described by the null hypothesis, given the sample size.

Because each sample is only a subset of the population scores, the sample does not exactly reproduce the population distribution. Thus, each set of sample scores contains some idiosyncratic variance, called "sampling error" variance, much like each person has idiosyncratic personality features. [Of course, sampling error variance should not be confused with either "measurement error" variance or "model specification" error variance (sometimes modeled as the "within" or "residual" sum of squares in univariate analyses) (Thompson, 1998a).] Of course, like people, sampling distributions may differ in how much idiosyncratic "flukiness" they each contain.

Statistical tests evaluate the probability that the deviation of the sample statistics from the assumed population parameters is due to sampling error. That is, statistical tests evaluate whether random sampling from the population may explain the deviations of the sample statistics from the hypothesized population parameters.

However, very few researchers employ random samples from the population. Rokeach (1973) was an exception; being a different person living in a different era, he was able to hire the Gallup polling organization to provide a representative national sample for his inquiry. But in the social sciences fewer than 5% of studies are based on random samples (Ludbrook & Dudley, 1998).

On the basis that most researchers do not have random samples from the population, some (cf. Shaver, 1993) have argued that statistical significance tests should almost never be used. However, most researchers presume that statistical tests may be reasonable if there are grounds to believe that the score sample of convenience is expected to be reasonably representative of a population.

In order to evaluate the probability that the sample scores came from a population of scores described exactly by the null hypothesis, given the sample size, researchers typically invoke the *sampling distribution*. The sampling distribution does <u>not</u> consist of scores (except when the sample size is one). Rather, the sampling distribution consists of estimated <u>parameters</u>, each computed for samples of exactly size <u>n</u>, so as to model the influences of random sampling error on the statistics estimating the population parameters, given the sample size.

This sampling distribution is then used to estimate the probability of the observed sample statistic(s) occurring due to sampling error. For example, we might take the population to be infinitely many IQ scores normally distributed with a mean, median and mode of 100 and a standard deviation of 15. Perhaps we have drawn a sample of 10 people, and compute the sample median (not all hypotheses have to be about means!) to be 110. We wish to know whether our statistic or one higher is unlikely, assuming the sample came from the posited population.

We can make this determination by drawing all possible samples of size 10 from the population, computing the median of each sample, and then creating the distribution of these statistics (i.e., the sampling distribution). We then examine the sampling distribution, and locate the value of 110. Perhaps only 2% of the sample statistics in the sampling distribution are 110 or higher. This suggests to us that our observed sample median of 110 is relatively unlikely to have come from the hypothesized population.

The number of samples drawn for the sampling distribution from a given population is a function of the population size, and the sample size. The number of such different sets of population cases for a population of size N and a sample of size n equals:

$$M = \frac{N!}{n!\ (N - n)!}.$$

Clearly, if the population size is infinite (or even only large), deriving all possible estimates becomes unmanageable. In such cases the sampling distribution may be theoretically (i.e., mathematically) estimated, rather than actually observed. Sometimes, rather than estimating the sampling distribution, estimating an analog of the sampling distribution, called a "test distribution" (e.g., F, t, χ^2) may be more manageable.

Heuristic Example for a Finite Population Case

Table 5 presents a finite population of scores for N=20 people. Presume that we wish to evaluate a sample mean for n=3 people. If we know (or presume) the population, we can derive the sampling distribution (or the test distribution) for this problem, so that we can then evaluate the probability that the sample statistic of interest came from the assumed population.

INSERT TABLE 5 ABOUT HERE.

Note that we are ultimately inferring the probability of the sample statistic, and not of the population parameter(s). Remember also that some specific population must be presumed, or infinitely many sampling distributions (and consequently infinitely $p_{CALCULATED}$ values) are plausible, and the solution becomes indeterminate.

Here the problem is manageable, given the relatively small population and samples sizes. The number of statistics creating this sampling distribution is

$$M = \frac{N!}{n!\ (N - n)!}$$

$$\frac{20!}{3!\ (20 - 3\ )!}$$

$$\frac{20!}{3!\ (17)!}$$

$$\frac{20x19x18x17x16x15x14x13x12x11x10x9x8x7x6x5x4x3x2}{3\ x\ 2\ x\ (17x16x15x14x13x12x11x10x9x8x7x6x5x4x3x2)}$$

$$\frac{2.433E+18}{6\quad x\quad 3.557E+14}$$

$$\frac{2.433E+18}{2.134E+15}$$

$$= 1,140.$$

Table 6 presents the first 85 and the last 10 potential samples. [The full sampling distribution takes 25 pages to present, and so is not presented here in its entirety.]

INSERT TABLE 6 ABOUT HERE.

Figure 6 presents the full sampling distribution of 1,140 estimates of the mean based on samples of size n=3 from the Table 5 population of N=20 scores. Figure 7 presents the analog of a test statistic distribution (i.e., the sampling distribution in standardized form).

INSERT FIGURES 6 AND 7 ABOUT HERE.

If we had a sample of size n=3, and had some reason to believe and wished to evaluate the probability that the sample with a mean of M = 524.0 came from the Table 5 population of N=20 scores, we could use the Figure 6 sampling distribution to do so. Statistic means (i.e., sample means) this large or larger occur about 25% of the time due to sampling error.

In practice researchers most frequently use sampling distributions of test statistics (e.g., F, t, χ^2), rather than the sampling distributions of sample statistics, to evaluate sample results. This is typical because the sampling distributions for many sample statistics change for every study variation (e.g., changes for different statistics, changes for each different sample size for even for a given statistic). Sampling distributions of test statistics (e.g., distributions of sample means each divided by the population SD) are more general or invariant over these changes, and thus, once they are estimated, can be used with greater regularity than the related sampling distributions for statistics.

The problem is that the applicability and generalizability of test distributions tend to be based on fairly strict assumptions (e.g., equal variances of outcome variable scores across all groups in ANOVA). Furthermore, test statistics have only been developed for a limited range of classical test statistics. For example, test distributions have *not* been developed for some "modern" statistics.

"Modern" Statistics

All "classical" statistics are centered about the arithmetic mean, M. For example, the standard deviation (SD), the coefficient of skewness (S), and the coefficient of kurtosis (K) are all moments about the mean, respectively:

$$SD_X = ((\Sigma\ (X_i - M_X)^2)\ /\ (n-1))^{.5} = ((\Sigma\ x_i^2)\ /\ (n-1))^{.5};$$

Coefficient of Skewness$_X$ (S_X) $= (\Sigma\ [(X_i-M_X)/SD_X]^3)\ /\ n$; and

Coefficient of Kurtosis$_X$ (K_X) $= ((\Sigma\ [(X_i-M_X)/SD_X]^4)\ /\ n) - 3$.

Similarly, the Pearson product-moment correlation invokes deviations from the means of the two variables being correlated:

$$r_{XY} = \frac{(\Sigma\ (X_i - M_X)(Y_i - M_Y))\ /\ n-1}{(SD_X * SD_Y)}.$$

The problem with "classical" statistics invoking the mean is that these estimates are notoriously influenced by atypical scores (outliers), partly because the mean itself is differentially influenced by outliers. Table 7 presents a heuristic data set that can be used to illustrate both these dynamics and two alternative "modern" statistics that can be employed to mitigate these problems.

INSERT TABLE 7 ABOUT HERE.

Wilcox (1997) presents an elaboration of some "modern" statistics choices. A shorter accessible treatment is provided by Wilcox (1998). Also see Keselman, Kowalchuk, and Lix (1998) and Keselman, Lix and Kowalchuk (1998).

The variable $\underline{X}$ in Table 7 is somewhat positively skewed ($\underline{S}_X$ = 2.40), as reflected by the fact that the mean ($\underline{M}_X$ = 500.00) is to the right of the median ($\underline{Md}_X$ = 461.00). One "modern" method "winsorizes" (à la statistician Charles Winsor) the score distribution by substituting less extreme values in the distribution for more extreme values. In this example, the 4th score (i.e., 433) is substituted for scores 1 through 3, and in the other tail the 17th score (i.e., 560) is substituted for scores 18 through 20. Note that the mean of this distribution, $\underline{M}_{X'}$ = 480.10, is less extreme than the original value (i.e., $\underline{M}_X$ = 500.00).

Another "modern" alternative "trims" the more extreme scores, and then computes a "trimmed" mean. In this example, .15 of the distribution is trimmed from each tail. The resulting mean, $\underline{M}_{X-}$ = 473.07, is closer to the median of the distribution, which has remained 461.00.

Some "classical" statistics can also be framed as "modern." For example, the interquartile range (75th %ile - 25th %ile) might be thought of as a "trimmed" range.

In theory, "modern" statistics may generate more replicable characterizations of data, because at least in some respects the influence of more extreme scores, which are less likely to be drawn in future samples from the tails of a non-uniform (non-rectangular or non-flat) population distribution, has been minimized. However, "modern" statistics have not been widely employed in contemporary research, primarily because generally-applicable test distributions are often not available for such statistics.

Traditionally, the tail of statistical significance testing has wagged the dog of characterizing our data in the most replicable manner. However, the "bootstrap" may provide a vehicle for statistically testing, or otherwise exploring, "modern" statistics.

Univariate Bootstrap Heuristic Example

The *bootstrap* logic has been elaborated by various methodologists, but much of this development has been due to Efron and his colleagues (cf. Efron, 1979). As explained elsewhere,

> Conceptually, these methods involve copying the data set on top of itself again and again infinitely many times to thus create an infinitely large "mega" data set (what's actually done is resampling from the original data set *with replacement*). Then hundreds or thousands of different samples [each of size n] are drawn from the "mega" file, and results [i.e., the statistics of interest] are computed separately for each sample and then averaged [and characterized

in various ways]. (Thompson, 1993b, p. 369)

Table 8 presents a heuristic data set to make concrete selected aspects of bootstrap analysis. The example involves the numbers of churches and murders in 45 cities. These two variables are highly correlated. [The illustration makes clear the folly of inferring causal relationships, even from a "causal modeling" SEM analysis, if the model is not exactly correctly "specified" (cf. Thompson, 1998a).] The statistic examined here is the bivariate product-moment correlation coefficient. This statistic is "univariate" in the sense that only a single dependent/outcome variable is involved.

INSERT TABLE 8 ABOUT HERE.

Figure 8 presents a scattergram portraying the linear relationship between the two measured/observed variables. For the heuristic data, r equals .779.

INSERT FIGURE 8 ABOUT HERE.

In this example 1,000 resamples of the rows of the Table 8 data were drawn, each of size n=45, so as to model the sampling error influences in the actual data set. In each "resample," because sampling from the Table 8 data was done "with replacement," a given row of the data may have been sampled multiple times, while another row of scores may not have been drawn at all. For this analysis the bootstrap software developed by Lunneborg (1987) was used. Table 9 presents some of the 1,000 bootstrapped estimates of r.

INSERT TABLE 9 ABOUT HERE.

Figure 9 presents a graphic representation of the bootstrap-estimated sampling distribution for this case. Because $\underline{r}$, although a characterization of linear relation, is not itself linear (i.e., $\underline{r}$=1.00 is not twice $\underline{r}$=.50), Fisher's $\underline{r}$-to-Z transformations of the 1,000 resampled $\underline{r}$ values were also computed as:

$\underline{r}$-to-Z = .5 (ln [(1 + $\underline{r}$)/(1 - $\underline{r}$)] (Hays, 1981, p. 465).

In SPSS this could be computed as:

```
compute r_to_z=.5 * ln ((1 + r)/(1 - r)).
```

Figure 10 presents the bootstrap-estimated sampling distribution for these values.

INSERT FIGURES 9 AND 10 ABOUT HERE.

Descriptive vs. Inferential Uses of the Bootstrap

The bootstrap *can* be used to test statistical significance. For example, the bootstrap can be used to estimate, through Monte Carlo simulation, sampling distributions when theoretical distributions (e.g., test distributions) are not known for some problems (e.g., "modern" statistics).

The standard deviation of the bootstrap-estimated sampling distribution characterizes the variability of the statistics estimating given population parameters. The standard deviation of the sampling distribution is called the "standard error of the estimate" (e.g., the standard error of the mean, $\underline{SE}_M$). [The decision to call this standard deviation the "standard error," so as to confuse the graduate students into not realizing that $\underline{SE}$ is an SD, was taken decades ago at an annual methodologists' coven--in the coven priority is typically afforded to most confusing the students regarding the most important concepts.] The SE of a statistic characterizes the precision or variability of the estimate.

The ratio of the statistic estimating a parameter to the SE of that estimate is a very important idea in statistics, and thus is called by various names, such as "t," "Wald statistic," and "critical ratio" (so as to confuse the students regarding an important concept). If the statistic is large, but the SE is even larger, a researcher may elect not to vest much confidence in the estimate. Conversely, even if a statistic is small (i.e., near zero), if the SE of the statistic is very, very small, the researcher may deem the estimate reasonably precise.

In classical statistics researchers typically estimate the SE as part of statistical testing by invoking numerous assumptions about the population and the sampling distribution (e.g., normality of the sampling distribution). Such SE estimates are theoretical.

The SD of the bootstrapped sampling distribution, on the other hand, is an empirical estimate of the sampling distribution's variability. This estimate does not require as many assumptions.

Table 10 presents selected percentiles for two bootstrapped r-to-z sampling distributions for the Table 8 data, one involving 100 resamples, and one involving 1,000 resamples. Notice that percentiles near the means or the medians of the two distributions tend to be closer than the values in the tails, and here especially in the left tail (small z values) where there are fewer values, because the distribution is skewed left. This purely heuristic comparison makes an extremely important conceptual point that clearly distinguishes inferential versus descriptive applications of the bootstrap.

INSERT TABLE 10 ABOUT HERE.

When we employ the bootstrap for inferential purposes (i.e., to estimate the probability of the sample statistics), focus shifts to the extreme tails of the distributions, where the less likely (and less frequent) statistics are located, because we typically invoke small values of p in statistical tests. These are exactly the locations where the estimated distribution densities are most unstable, because there are relatively few scores here (presuming the sampling distribution does not have an extraordinarily small SE). Thus, when we invoke the bootstrap to conduct statistical significance tests, extremely large numbers of resamples are required (e.g., 2,000, 5,000).

However, when our application is descriptive, we are primarily interested in the mean (or median) statistic and the SD/SE from the sampling distribution. These values are less dependent on large numbers of resamples. This is said not to discourage large numbers of resamples (which are essentially free to use, given modern microcomputers), but is noted instead to emphasize these two very distinct uses of the bootstrap.

The descriptive focus is appropriate. We hope to avoid obtaining results that no one else can replicate (partly because we are good scientists searching for generalizable results, and partly simply because we do not wish to be embarrassed by discovering the social sciences equivalent of cold fusion). The challenge is obtaining results that reproduce over the wide range of idiosyncracies of human personality.

The descriptive use of the bootstrap provides some evidence, short of a real (and preferred) "external" replication (cf. Thompson, 1996) of our study, that results may generalize. As noted elsewhere,

> If the mean estimate [in the estimated sampling distribution] is like our sample estimate, and the standard deviation of estimates from the resampling is small, then we have some indication that the result is stable over many different configurations of subjects. (Thompson, 1993b, p. 373)

Multivariate Bootstrap Heuristic Example

The bootstrap can also be generalized to multivariate cases (e.g., Thompson, 1988b, 1992a, 1995a). The barrier to this application is that a given multivariate "factor" (also called "equation," "function," or "rule," for reasons that are, by now, obvious) may be manifested in different locations.

For example, perhaps a measurement of androgyny purports to measure two factors: masculine and feminine. In one resample masculine may be the first factor, while in the second resample masculine might be the second factor. In most applications we have no particular theoretical expectation that "factors" ("functions," etc.) will always replicate in a given order. However, if we average and otherwise characterize statistics across resamples without initially locating given constructs in the same locations, we will be pooling apples, oranges, and tangerines, and merely be creating a mess.

This barrier to the multivariate use of the bootstrap can be resolved by using Procrustean methods to rotate all "factors" into a single, common factor space prior to characterizing the results across the resamples. A brief example may be useful in communicating the procedure.

Figure 11 presents DDA/MANOVA results from an analysis of Sir Ronald Fisher's (1936) classic data for iris flowers. Here the bootstrap was conducted using my DISCSTRA program (Thompson, 1992a) to conduct 2,000 resamples.

INSERT FIGURE 11 ABOUT HERE.

Figure 12 presents a partial listing of the resampling of _n_=150 rows of data (i.e., the resample size exactly matches the original samples size). Notice in Figure 12 that case #27 was selected at least twice as part of the first resample.

INSERT FIGURE 12 ABOUT HERE.

First 13 presents selected results for both the first and the last resamples. Notice that the function coefficients are first rotated to best fit position with a common designated target matrix, and then the structure coefficients are computed using these rotated results. [Here the rotations made few differences, because the functions by happenstance already fairly closely matched the target matrix--here the function coefficients from the original sample.]

INSERT FIGURE 13 ABOUT HERE.

Figure 14 presents an abridged map of participant selection across the 2,000 resamples. We can see that the 150 flowers were each selected approximately 2,000 times, as expected if the random selection with replacement is truly random.

INSERT FIGURE 14 ABOUT HERE.

Figure 15 presents a summary of the bootstrap DDA results. For example, the mean statistic across 2,000 resample is computed along with the empirically-estimated standard error of each statistic. As generally occurs, SE's tend to be smaller for statistics that deviate most from zero; these coefficients tend to reflect real (non-sampling error variance) dynamics within the data, and therefore tend to re-occur across samples.

INSERT FIGURE 15 ABOUT HERE.

However, notice in Figure 15 that the SE's for the standardized function coefficients on Function I for variables X2 and X4 were both essentially .40, even though the mean estimates of the two coefficients appear to be markedly different (i.e., ¦1.6¦ and ¦2.9¦). In a theoretically-grounded estimate, for a given n and a given population estimate, the SE will be identical. But bootstrap methods do not require the sometimes unrealistic assumption that related coefficients even in a given analysis with a common fixed n have the same sampling distributions.

Clarification and an Important Caveat

The bootstrap methods modeled here presume that the sample size is somewhat large (i.e., more than 20 to 40). In these cases the bootstrap invokes resampling with replacement. For small samples other methods are employed.

It is also important to emphasize that "bootstrap methods do not magically take us beyond the limits of our data" (Thompson, 1993b, p. 373). For example, the bootstrap cannot make an unrepresentative sample representative. And the bootstrap cannot make a quasi-experiment with intact groups mimic results for a true experiment in which random assignment is invoked. The bootstrap cannot make data from a correlational (i.e., non-experimental) design yield unequivocal causal conclusions.

Thus, Lunneborg (1999) makes very clear and careful distinctions between bootstrap applications that may support either (a) population inference (i.e., the study design invoked random sampling), or (b) evaluation of how "local" a causal inference may be (i.e., the study design invoked random assignment to experimental groups, but not random selection), or (c) evaluation of how "local" non-causal descriptions may be (i.e., the design invoked neither random sampling nor random assignment). Lunneborg (1999) quite rightly emphasizes how critical it is to match study design/purposes and the bootstrap modeling procedures.

The bootstrap and related "internal" replicability analyses are not magical. Nevertheless, these methods can be useful because

> the methods combine the subjects in hand in [numerous] different ways to determine whether

> results are stable across sample variations, i.e., across the idiosyncracies of individuals which make generalization in social science so challenging. (Thompson, 1996, p. 29)

Effect Sizes

As noted previously, $\underline{p}_{\text{CALCULATED}}$ values are not suitable indices of effect, "because both [types of $\underline{p}$ values] *depend on sample size*" (APA, 1994, p. 18, emphasis added). Furthermore, unlikely events are not intrinsically noteworthy (see Shaver's (1985) classic example). Consequently, the APA publication manual now "encourages" (p. 18) authors to report effect sizes.

Unfortunately, a growing corpus of empirical studies of published articles portrays a consensual view that merely "encouraging" effect size reporting (APA, 1994) has not appreciably affected actual reporting practices (e.g., Keselman et al., 1998; Kirk, 1996; Lance & Vacha-Haase, 1998; Nilsson & Vacha-Haase, 1998; Reetz & Vacha-Haase, 1998; Snyder & Thompson, 1998; Thompson, 1999b; Thompson & Snyder, 1997, 1998; Vacha-Haase & Ness, 1999; Vacha-Haase & Nilsson, 1998). Table 11 summarizes 11 empirical studies of recent effect size reporting practices in 23 journals.

INSERT TABLE 11 ABOUT HERE.

Although some of the Table 11 results appear to be more favorable than others, it is important to note that in some of the 11 studies' effect sizes were counted as being reported even if the relevant results were not interpreted (e.g., an $\underline{r}^2$ was reported but not interpreted as being big or small, or noteworthy or not). This dynamic is dramatically illustrated in the Keselman et al. (1998) results, because the reported results involved an exclusive focus on between-subjects OVA designs, and thus there were no spurious counts of incidental variance-accounted-for statistic reports. Here Keselman et al. (1998) concluded that, "as anticipated, effect sizes were almost never reported along with p-values" (p. 358).

If the baseline expectation is that effect should be reported in 100% of quantitative studies (mine is), the Table 11 results are disheartening. Elsewhere I have presented various reasons why I anticipate that the current APA (1994, p. 18) "encouragement" will remain largely ineffective. I have noted that an "encouragement" is so vague as to be unenforceable (Thompson, in press-b). I have also observed that only "encouraging" effect size reporting:

> presents a self-canceling mixed-message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, in press-b)

Two Heuristic Hypothetical Literatures

Two heuristic hypothetical literatures can be presented to illustrate the deleterious impacts of contemporary traditions. Here, results are reported for both statistical tests and effect sizes.

Twenty "TinkieWinkie" Studies. First, presume that a televangalist suddenly denounces a hypothetical childrens' television character, "TinkieWinkie," based on a claim that the character intrinsically by appearance and behavior incites moral depravity in 4 year olds.

This claim immediately incites inquiries by 20 research teams, each working independently without knowledge of each others' results. These researchers conduct experiments comparing the differential effects of "The TinkieWinkie Show" against those of "Sesame Street," or "Mr. Rogers," or both.

This work results in the nascent new literature presented in Table 12. The eta^2 effect sizes from the 20 (10 two-level one-way and 10 three-level one-way) ANOVAs range from 1.2% to 9.9% ($\underline{M}_{sq\ eta}$=3.00%; $\underline{SD}_{sq\ eta}$=2.0%) as regards moral depravity being induced by "The TinkieWinkie Show." However, as reported in Table 12, only 1 of the 20 studies results in a statistically significant effect.

INSERT TABLE 12 ABOUT HERE.

The 19 research teams finding no statistically significant differences in the treatment effects on the moral depravity of 4 year olds obtained effect sizes ranging from eta^2=1.2% to eta^2=4.8%. Unfortunately, these 19 research teams are acutely aware of how non-statistically significant findings are valued within the profession.

They are acutely aware, for example, that revised versions of published articles were rated more highly by counseling practitioners if the revisions reported statistically significant findings than if they reported statistically nonsignificant findings (Cohen, 1979). The research teams are also acutely aware of Atkinson, Furlong and Wampold's (1982) study in which

> 101 consulting editors of the Journal of Counseling Psychology and the Journal of Consulting and Clinical Practice were asked to evaluate three versions, differing only with regard to level of statistical significance, of a research manuscript. The statistically nonsignificant and approach significance versions were more than three times as likely to be recommended for rejection than was the statistically significant version. (p. 189)

Indeed, Greenwald (1975) conducted a study of 48 authors and 47 reviewers for the Journal of Personality and Social Psychology and reported a

> 0.49 (± .06) probability of submitting a rejection of the null hypothesis for publication (Question 4a) compared to the low probability of 0.06 (± .03) for submitting a nonrejection of the null hypothesis for publication (Question 5a). A secondary bias is apparent [as well] *in the probability of continuing with a problem [in future inquiry]*. (p. 5, emphasis added)

This is the well known "file drawer problem" (Rosenthal, 1979). In the present instance, some of the 19 research teams failing to reject the null hypothesis decide not to even submit their work, while the remaining teams have their reports rejected for publication. Perhaps these researchers were socialized by a previous version of the APA publication manual, which noted that:

> Even when the theoretical basis for the prediction

> is clear and defensible, the burden of methodological precision falls heavily on the investigator who reports negative results. (APA, 1974, p. 21)

Here only the one statistically significant result is published; everyone remains happily oblivious to the overarching substance of the literature in its entirety.

The problem is that setting a low alpha only means that the probability of a Type I error will be small on the average. In the literature as a whole, some unlikely Type I errors are still inevitable. These will be afforded priority for publication. Yet publishing replication disconfirmations of these Type I errors will be discouraged normatively. Greenwald (1975, pp. 13-15) cites the expected actual examples of such epidemics. In short, contemporary practice as regards statistical tests actively discourages some forms of replication, or at least discourages disconfirming replications being published.

Twenty Cancer Treatment Studies. Here researchers learn of a new theory that a newly synthesized protein regulates the growth of blood supply to cancer tumors. It is theorized that the protein might be used to prevent new blood supplies from flowing to new tumors, or even that the protein might be used to reduce existing blood flow to tumors and thus lead to cancer destruction. The protein is synthesized.

Unfortunately, given the newness of the theory and the absence of previous related empirical studies upon which to ground power analyses for their new studies, the 20 research teams institute inquiries that are slightly under-powered. The results from these 20 experiments are presented in Table 13.

INSERT TABLE 13 ABOUT HERE.

Here all 20 studies yield $\underline{p}_{\text{CALCULATED}}$ values of roughly .06 (range = .0598 to .0605). As reported in Table 13, the effect sizes range from 15.1% to 62.8%. In the present scenario, only a few of the reports are submitted for publication, and none are published.

Yet, these inquiries yielded effect sizes ranging from eta^2=15.1%, which Cohen (1988, pp. 26-27) characterized as "large," at least as regards result typicality, up to eta^2=62.8%. And a life-saving outcome variable is being measured! At the individual study level, perhaps each research team has decided that $\underline{p}$ values evaluate result replicability, and remain oblivious to the uniformity of efficacy findings across the literature.

Some researchers remain devoted to statistical tests, because of their professed dedication to reporting only replicable results, and because they erroneously believe that statistical significance evaluates result replicability (Cohen, 1994). In summary, *it would be the abject height of irony if, out of devotion to replication, we continued to worship at the tabernacle of statistical significance testing, and at the same time we declined to (a) formulate our hypotheses by explicit consultation of the effect sizes reported in previous studies and (b) explicitly interpret our obtained effect sizes in relation to those reported in related previous inquiries.*

An Effect Size Primer

Given the central role that effect sizes should play with quantitative studies, at least a brief review of the available choices is warranted here. Very good treatments are also available from Kirk (1996), Rosenthal (1994), and Snyder and Lawson (1993).

There are dozens of effect size estimates, and no single one-size-fits-all choice. The effect sizes can be divided into two major classes: (a) standardized differences and (b) variance-accounted-for measures of strength of association. [Kirk (1996) identifies a third, "miscellaneous" category, and also summarizes some of these choices.]

Standardized differences. In experimental studies, and especially studies with only two groups where the mean is of primary interest, the differences in means can be "standardized" by dividing the difference by some estimate of the population parameter score σ. For example, in his seminal work on meta-analysis, Glass (cf. 1976) proposed that the difference in the two means could be divided by the *control group* standard deviation to estimate Δ.

Glass presumed that the control group standard deviation is the best estimate of σ. This is reasonable particularly if the control group received no treatment, or a placebo treatment. For example, for the Table 2 variable, X, if the second of the two groups was taken as the control group,

$$\Delta_X = (12.50 - 11.50) / 7.68 = .130.$$

In this estimation the variance (see Table 2 note) is computed by dividing the sum of squares by n-1.

However, others have taken the view that the most accurate standardization can be realized by use of a *"pooled"* (across groups) estimate of the population standard deviation. Hedges (1981) advocated computation of g using the standard deviation computed as the square root of a pooled variance based on division of the sum of squares by n-1. For the Table 2 variable, X,

$$g_X = (12.50 - 11.50) / 7.49 = .134.$$

Cohen (1969) argued for the use of d, which divides the mean difference by a *"pooled"* standard deviation computed as the square root of a pooled variance based on division of the sum of squares by n. For the Table 2 variable, X,

$$d_X = (12.50 - 11.50) / 7.30 = .137.$$

As regards these choices, there is (as usual) no one always right one-size-fits-all choice. The comment by Huberty and Morris (1988, p. 573) is worth remembering generically: "As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness!"

In some studies the control group standard deviation provides the most reasonable standardization, while in others a "pooling" mechanism may be preferred. For example, an intervention may itself change score variability, and in these cases Glass's Δ may be preferred. But otherwise the "pooled" value may provide the more statistically precise estimate.

As regards correction for statistical bias by division by n-1 versus n, of course the competitive differences here are a function of the value of n. As n gets larger, it makes less difference which choice is made. This division is equivalent to multiplication by 1 / the divisor. Consider the differential impacts on estimates derived using the following selected choices of divisors.

n	1/Divisor	n-1	1/Divisor	Difference
10	.1000	9	.111111	.011111
100	.0100	99	.010101	.000101
1000	.0010	999	.001001	.000001
10000	.0001	9999	.000100010	.00000001

Variance-accounted-for. Given the omnipresence of the General Linear Model, all analyses are correlational (cf. Thompson, 1998a), and (as noted previously) an r^2 effect size (e.g., eta^2, $\underline{R}^2$, omega^2 [ω^2; Hays, 1981], adjusted $\underline{R}^2$) can be computed in all studies. Generically, in univariate analyses "uncorrected" variance-accounted-for effect sizes (e.g., eta^2, $\underline{R}^2$) can be computed by dividing the sum of squares "explained" ("between," "model," "regression") by the sum of squares of the outcome variable (i.e., the sum of squares "total"). For example, in the Figure 3 results, the univariate eta^2 effect sizes were both computed to be 0.469% (e.g., 5.0 / [5.0 + 1061.0] = 5.0 / 1065.0 = .00469).

In multivariate analysis, one estimate of eta^2 can be computed as 1 - lambda (λ). For example, for the Figure 3 results, the multivariate eta^2 effect size was computed as (1 - .37500) equals .625.

Correcting for score measurement unreliability. It is well known that score unreliability tends to attenuate $\underline{r}$ values (cf. Walsh, 1996). Thus, some (e.g., Hunter & Schmidt, 1990) have recommended that effect sizes be estimated incorporating statistical corrections for measurement error. However, such corrections must be used with caution, because any error in estimating the reliability will considerably distort the effect sizes (cf. Rosenthal, 1991).

Because scores (not tests) are reliable, reliability coefficients fluctuate from administration to administration (Reinhardt, 1996). In a given empirical study, the reliability for the data in hand may be used for such corrections. In other cases, more confidence may be vested in these corrections if the reliability estimates employed are based on the important meta-analytic "reliability generalization" method proposed by Vacha-Haase (1998).

"Corrected" vs. "uncorrected" variance-accounted-for estimates. "Classical" statistical methods (e.g., ANOVA, regression, DDA) use the statistical theory called "ordinary least squares." This theory optimizes the fit of the synthetic/latent variables (e.g., $\hat{Y}$) to the observed/measured outcome/response variables (e.g., $\underline{Y}$) in the sample data, and capitalizes on all the variance present in the observed sample scores, including the "sampling error variance" that it is idiosyncratic to the particular sample. Because sampling error variance is unique to a given sample (i.e., each sample has its own sampling error variance), "uncorrected" variance-accounted-for effect sizes somewhat overestimate the effects that would be replicated by applying the same weights (e.g., regression beta weights) in either (a) the population or (b) a different sample.

However, statistical theory (or the descriptive bootstrap) can be invoked to estimate the extent of overestimation (i.e., positive bias) in the variance-accounted-for effect size estimate. [Note that "corrected" estimates are always less than or equal to "uncorrected" values.] The difference between the "uncorrected" and "corrected" variance-accounted-for effect sizes is called "shrinkage."

For example, for regression the "corrected" effect size "adjusted $\underline{R}^2$" is routinely provided by most statistical packages. This correction is due to Ezekiel (1930), although the formula is often incorrectly attributed to Wherry (Kromrey & Hines, 1996):

$$1 - ((\underline{n} - 1) / (\underline{n} - \underline{v} - 1)) \times (1 - \underline{R}^2),$$

where $\underline{n}$ is the sample size and $\underline{v}$ is the number of predictor variables. The formula can be equivalently expressed as:

$$\underline{R}^2 - ((1 - \underline{R}^2) \times (\underline{v} / (\underline{n} - \underline{v} -1))).$$

In the ANOVA case, the analogous η^2 can be computed using the formula due to Hays (1981, p. 349):

$$(SS_{BETWEEN} - (\underline{k} - 1) \times MS_{WITHIN}) / (SS_{TOTAL} + MS_{WITHIN}),$$

where $\underline{k}$ is the number of groups.

In the multivariate case, a multivariate omega^2 due to Tatsuoka (1973a) can be used as "corrected" effect estimate. Of course, using univariate effect sizes to characterize multivariate results would be just as wrong-headed as using ANOVA methods *post hoc* to MANOVA. As Snyder and Lawson (1993) perceptively noted, "researchers asking multivariate questions will need to use magnitude-of-effect indices that are consistent with their multivariate view of the research problem" (p. 341).

Although "uncorrected" effects for a sample are larger than the "corrected" effects estimated for the population, the "corrected" estimates for the population effect (e.g., omega^2) tend in turn to be larger than the "corrected" estimates for a future sample (e.g., Herzberg, 1969; Lord, 1950). As Snyder and Lawson

(1993) explained, "the reason why estimates for future samples result in the most *shrinkage* is that these statistical corrections must adjust for the sampling error present in *both* the given present study and some future study" (p. 340, emphasis in original).

It should also be noted that variance-accounted-for effect sizes can be negative, notwithstanding the fact that a squared-metric statistic is being estimated. This was seen in some of the omega^2 values reported in Table 12. Dramatic amounts of shrinkage, especially to negative variance-accounted-for values, suggest a somewhat dire research experience. Thus, I was somewhat distressed to see a local dissertation in which $\underline{R}^2$=44.6% shrunk to 0.45%, and yet it was claimed that still "it may be possible to generalize prediction in a referred population" (Thompson, 1994a, p. 12).

Factors that inflate sampling error variance. Understanding what design features generate sampling error variance can facilitate more thoughtful design formulation, and thus has some value in its own right. Sampling error variance is *greater* when:

(a) *sample size* is *smaller;*

(b) the number of *measured variables* is *greater;* and

(c) the *population effect size* (i.e., parameter) is *smaller.*

The deleterious effects of small sample size are obvious. When we sample, there is more likelihood of "flukie" characterizations of the population with smaller samples, and the relative influence of anomalous scores (i.e., outliers) is greater in smaller samples, at least if we use "classical" as against "modern" statistics.

Table 14 illustrates these variations as a function of different sample sizes for regression analyses each involving 3 predictor variables and presumed population parameter R^2 equal to 50%. These results illustrate that the sampling error due to sample size is not a monotonic (i.e., constant linear) function of sample size changes. For example, when sample size changes from n=10 to n=20, the shrinkage changes from 25.00% (R^2=50% - R^2*=25.00%) to 9.73% (R^2=50% - R^2*=40.63%). But even more than doubling sample size from n=20 to n=45 changes shrinkage only from 9.73% (R^2=50% - R^2*=40.63%) to 3.66% (R^2=50% - R^2*=46.34%).

INSERT TABLE 14 ABOUT HERE.

The influence of the *number of measured variables* is also fairly straightforward. The more variables we sample the greater is the likelihood that an anomalous score will be incorporated in the sample data.

The common language describing a person as an "outlier" should *not* be erroneously interpreted to mean either (a) that a given person is an outlier on all variables or (b) that a given score is an outlier as regards all statistics (e.g., on the mean versus the correlation). For example, for the following data Amanda's score may be outlying as regards M_Y, but not as regards r_{XY} (which here equal +1; see Walsh, 1996).

Person	X_i	Y_i
Kevin	1	2
Jason	2	4
Sherry	3	6
Amanda	48	96

Again, as reported in Table 14, the influence of the number of measured variables on shrinkage is not monotonic.

Less obvious is why the estimated population parameter effect size (i.e., the estimate based on the sample statistic) impacts shrinkage. The easiest way to understand this is to conceptualize the population for a Pearson product-moment study. Let's say the population squared correlation is +1. In this instance, even ridiculously small samples of any 2 or 3 or 4 pairs of scores will invariably yield a sample $\underline{r}^2$ of 100% (as long as both $\underline{X}$ and $\underline{Y}$ as sampled are variables, and therefore $\underline{r}$ is "defined," in that illegal division is not required by the formula $\underline{r} = \underline{COV}_{XY}$ / [$\underline{SD}_X$ x $\underline{SD}_Y$]).

Again as suggested by the Table 14 examples, the influence of increased sample size on decreased shrinkage is not monotonic. [Thus, the use of a sample $\underline{r}$=.779 in the Table 8 heuristic data for the bootstrap example theoretically should have resulted in relatively little variation in sample estimates across resamples.]

Indeed, these three influences on sampling error must be considered as they simultaneously interact with each other. For example, as suggested by the previous discussion, the influence of sample size is an influence conditional on the estimated parameter effect size. Table 15 illustrates these interactions for examples all of which involve shrinkage of a 5% decrement downward from the original $\underline{R}^2$ value.

INSERT TABLE 15 ABOUT HERE.

Pros and cons of the effect size classes. It is not clear that researchers should uniformly prefer one effect index over another, or even one class of indices over the other. The standardized difference indices do have one considerable advantage: they tend to be readily comparable across studies because they are expressed "metric-free" (i.e., the division by *SD* removes the metric from the characterization).

However, variance-accounted-for effect sizes can be directly computed in all studies. Furthermore, the use of variance-accounted-for effect sizes has the considerable heuristic value of forcing researchers to recognize that all parametric methods are part of a single general linear model family (cf. Cohen, 1968; Knapp, 1978).

In any case, the two effect sizes can be re-expressed in terms of each other. Cohen (1988, p. 22) provided a general table for this purpose. A *d* can also be converted to an *r* using Cohen's (1988, p. 23) formula #2.2.6:

$$\begin{aligned}
\underline{r} &= \underline{d} \;/\; [(\underline{d}^2 + 4)^{.5}] \\
&= \underline{0.8} \;/\; [(0.8^2 + 4)^{.5}] \\
&= 0.8 \;/\; [(0.64 + 4)^{.5}] \\
&= 0.8 \;/\; [(\ 4.64\ )^{.5}] \\
&= 0.8 \;/\; 2.154 \\
&= \underline{0.371}\ .
\end{aligned}$$

An *r* can be converted to a *d* using Friedman's (1968, p. 246) formula #6:

$$\begin{aligned}
\underline{d} &= [2\ (\underline{r})] \;/\; [(1 - \underline{r}^2)^{.5}] \\
&= [2\ (\underline{0.371})] \;/\; [(1 - 0.371^2)^{.5}] \\
&= [2\ (0.371)] \;/\; [(1 - 0.1376)^{.5}] \\
&= [2\ (0.371)] \;/\; (0.8624)^{.5} \\
&= [2\ (0.371)] \;/\; 0.9286 \\
&= 0.742 \;/\; 0.9286 \\
&= \underline{0.799}\ .
\end{aligned}$$

Effect Size Interpretation. Schmidt and Hunter (1997) recently argued that "logic-based arguments [against statistical testing] seem to have had only a limited impact... [perhaps due to] the virtual brainwashing in significance testing that all of us have undergone" (pp. 38-39). They also spoke of a "psychology of addiction to significance testing" (Schmidt & Hunter, 1997, p. 49).

For too long researchers have used statistical significance tests in an illusory atavistic escape from the responsibility for defending the value of their results. Our p values were implicitly invoked as the universal coinage with which to argue result noteworthiness (and replicability). But as I have previously noted,

> Statistics can be employed to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an atavistic escape (à la Fromm's Escape from Freedom) from the existential human responsibility for making value judgments. If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating p's, and so p's cannot be blithely used to infer the value of research results. (Thompson, 1993b, p. 365)

The problem is that the normative traditions of contemporary social science have not yet evolved to accommodate personal values explication as part our work. As I have suggested elsewhere (Thompson, 1999a),

> Normative practices for evaluating such [values] assertions will have to evolve. Research results should not be published merely because the

> individual researcher thinks the results are noteworthy. By the same token, editors should not quash research reports merely because they find explicated values unappealing. These resolutions will have to be formulated in a spirit of reasoned comity. (p. 175)

In his seminal book on power analysis, Cohen (1969, 1988, pp. 24-27) suggested values for what he judged to be "low," "medium," and "large" effect sizes:

Characterization	*d*	r^2
"low"	.2	1.0%
"medium"	.5	5.9%
"large"	.8	13.8%

Cohen (1988) was characterizing what he regarded as the typicality of effect sizes across the broad published literature of the social sciences. However, some empirical studies suggest that Cohen's characterization of typicality is reasonably accurate (Glass, 1979; Olejnik, 1984).

However, as Cohen (1988) himself emphasized:

> The terms "small," "medium," and "large" are relative, not only to each other, but to the content area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation... In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions... in as diverse a field of inquiry as behavioral science... [This] common conventional frame of

> reference... is recommended for use *only when no better basis for estimating the ES index is available*. (p. 25, emphasis added)

If in evaluating effect size we apply Cohen's conventions (against his wishes) with the same rigidity with which we have traditionally applied the α=.05 statistical significance testing convention we will merely be being stupid in a new metric.

In defending our subjective judgments that an effect size is noteworthy in our personal value system, we must recognize that inherently any two researchers with individual values differences may reach different conclusions regarding the noteworthiness of the exact same effect even in the same study. And, of course, the same effect size in two different inquiries may differ radically in noteworthiness. Even small effects will be deemed noteworthy, if they are replicable, when inquiry is conducted as regards highly valued outcomes. Thus, Gage (1978) pointed out that even though the relationship between cigarette smoking and lung cancer is relatively "small" (i.e., $\underline{r}^2$ = 1% to 2%):

> Sometimes even very weak relationships can be important... [O]n the basis of such correlations, important public health policy has been made and millions of people have changed strong habits. (p. 21)

<u>Confidence Intervals for Effects</u>. It often is useful to present confidence intervals for effect sizes. For example, a series of confidence intervals across variables or studies can be conveyed in a concise and powerful graphic. Such intervals might

incorporate information regarding the theoretical or the empirical (i.e., bootstrap) estimates of effect variability across samples. However, as I have noted elsewhere,

> If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing. But if we interpret the confidence intervals in our study in the context of the intervals in all related previous studies, the true population parameters will eventually be estimated across studies, even if our prior expectations regarding the parameters are wildly wrong (Schmidt, 1996). (Thompson, 1998b, p. 799)

Conditions Necessary (and Sufficient) for Change

Criticisms of conventional statistical significance are not new (cf. Berkson, 1938; Boring, 1919), though the publication of such criticisms does appears to be escalating at an exponentially increasing rate (Anderson et al., 1999). Nearly 40 years ago Rozeboom (1960) observed that "the perceptual defenses of psychologists [and other researchers, too] are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

Table 16 summarizes some of the features of contemporary practice, the problems associated with these practices, and potential improvements in practice. The implementation of these "modern" inquiry methods would result in the more thoughtful specification of research hypotheses. The design of studies with more statistical power and precision would be more likely, because power analyses would be based on more informed and realistic effect size estimates as an effect literature matured (Rossi, 1997).

INSERT TABLE 16 ABOUT HERE.

Emphasizing effect size reporting would eventually facilitate the development of theories that support more specific expectations. Universal effect size reporting would facilitate improved meta-analyses of literature in which cumulated effects would not be based on as many strong assumptions that are probably somewhat infrequently met. Social science would finally become the business of identifying valuable effects that replicate under stated conditions; replication would no longer receive the hollow affection of the statistical significance test, and instead the replication of specific effects would be explicitly and directly addressed.

What are the conditions necessary and sufficient to persuade researchers to pay less attention to the likelihood of sample statistics, based on assumptions that "nil" null hypotheses are true in the population, and more attention to (a) effect sizes and (b) evidence of effect replicability? Certainly current doctoral curricula seem to have less and less space for quantitative training (Aiken et al., 1990). And too much instruction teaches analysis as the rote application of methods *sans* rationale (Thompson, 1998a). And many textbooks, too, are flawed (Carver, 1978; Cohen, 1994).

But improved textbooks will not alone provide the magic bullet leading to improved practice. The computation and interpretation of effect sizes are already emphasized in some texts (cf. Hays, 1981). For example, Loftus and Loftus (1982) in their book argued that "it is our judgment that accounting for variance is really much more meaningful than testing for [statistical] significance" (p. 499).

Editorial Policies

I believe that changes in journal editorial policies are the necessary (and sufficient) conditions to move the field. As Sedlmeier and Gigerenzer (1989) argued, "there is only one force that can effect a change, and that is the same force that helped institutionalize null hypothesis testing as the *sine qua non* for publication, namely, the editors of the major journals" (p. 315). Glantz (1980) agreed, noting that "The journals are the major force for quality control in scientific work" (p. 3). And as Kirk (1996) argued, changing requirements in journal editorial policies as regards effect size reporting "would cause a chain reaction: Statistics teachers would change their courses, textbook authors would revise their statistics books, and journal authors would modify their inference strategies" (p. 757).

Fortunately, some journal editors have elaborated policies "requiring" rather than merely "encouraging" (APA, 1994, p. 18) effect size reporting (cf. Heldref Foundation, 1997, pp. 95-96; Thompson, 1994b, p. 845). It is particularly noteworthy that editorial policies even at one APA journal now indicate that:

> If an author decides not to present an effect size estimate along with the outcome of a significance

> test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

Leadership from AERA

Professional disciplines, like glaciers, move slowly, but inexorably. The hallmark of a profession is standards of conduct. And, as Biesanz and Biesanz (1969) observed, "all members of the profession are considered colleagues, equals, who are expected to uphold the dignity and mystique of the profession in return for the protection of their colleagues" (p. 155). Especially in academic professions, there is some hesitance to change existing standards, or to impose more standards than seem necessary to realize common purposes.

As might be expected, given these considerations, in its long history AERA has been reticent to articulate standards for the conduct of educational inquiry. Most such expectations have been articulated only in conjunction with other organizations (e.g., AERA/APA/NCME, 1985). For example, AERA participated with 15 other organizations in the Joint Committee on Standards for Educational Evaluation's (1994) articulation of the program evaluation standards. These were the first-ever American National Standards Institute (ANSI)-approved standards for professional conduct. As ANSI-approved standards, these represent *de facto* THE American standards for program evaluation (cf. Sanders, 1994).

As Kaestle (1993) noted some years ago,

> ...[I]f education researchers could reverse their reputation for irrelevance, politicization, and disarray, however, they could rely on better support because most people, in the government and the public at large, believe that education is critically important. (pp. 30-31)

Some of the desirable movements of the field may be facilitated by the on-going work of the APA Task Force on Statistical Inference (Azar, 1997; Shea, 1996).

But AERA, too, could offer academic leadership. The children who are served by education need not wait for AERA to wait for APA to lead via continuing revisions of the APA publication manual. AERA, through the new Research Advisory Committee, and other AERA organs, might encourage the formulation of editorial policies that place less emphasis on statistical tests based on "nil" null hypotheses, and more emphasis on evaluating whether educational interventions and theories yield valued effect sizes that replicate under stated conditions.

It would be a gratifying experience to see our organization lead movement of the social sciences. Offering credible academic leadership might be one way that educators could confront the "awful reputation" (Kaestle, 1993) ascribed to our research. As I argued 3 years ago, if education "studies inform best practice in classrooms and other educational settings, the stakeholders in these locations certainly deserve better treatment from the [educational] research community via our analytic choices" (p. 29).

References

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Erlbaum.

Aiken, L.S., West, S.G., Sechrest, L., Reno, R.R., with Roediger, H.L., Scarr, S., Kazdin, A.E., & Sherman, S.J. (1990). The training in statistics, methodology, and measurement in psychology. American Psychologist, 45, 721-734.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: Author.

American Psychological Association. (1974). Publication manual of the American Psychological Association (2nd ed.). Washington, DC: Author.

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

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (1999). Null hypothesis testing in ecological studies: Problems, prevalence, and an alternative. Manuscript submitted for publication.

Anonymous. (1998). [Untitled letter]. In G. Saxe & A. Schoenfeld, Annual meeting 1999. Educational Researcher, 27(5), 41.

*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.

Atkinson, R.C., & Jackson, G.B. (Eds.). (1992). Research and education reform: Roles for the Office of Educational Research and Improvement. Washington, DC: National Academy of Sciences. (ERIC Document Reproduction Service No. ED 343 961)

Azar, B. (1997). APA task force urges a harder look at data. The APA Monitor, 28(3), 26.

Bagozzi, R.P., Fornell, C., & Larcker, D.F. (1981). Canonical correlation analysis as a special case of a structural relations model. Multivariate Behavioral Research, 16, 437-454.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Barnette, J.J., & McLean, J.E. (1998, November). Protected versus unprotected multiple comparison procedures. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526-536.

Biesanz, J., & Biesanz, M. (1969). Introduction to sociology. Englewood Cliffs, NJ: Prentice-Hall.

References designated with asterisks are empirical studies of research practices.

Borgen, F.H., & Seling, M.J. (1978). Uses of discriminant analysis following MANOVA: Multivariate statistics for multivariate purposes. Journal of Applied Psychology, 63, 689-697.

Boring, E.G. (1919). Mathematical vs. scientific importance. Psychological Bulletin, 16, 335-338.

Breunig, N.A. (1995, November). Understanding the sampling distribution and its use in testing statistical significance. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS. (ERIC Document Reproduction Service No. ED 393 939)

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

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

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, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

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

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

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.

Cortina, J.M., & Dunlap, W.P. (1997). Logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.

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

Davison, A.C., & Hinkley, D.V. (1997). Bootstrap methods and their applications. Cambridge: Cambridge University Press.

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

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

*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.

Efron, B., & Tibshirani, R.J. (1993). An introduction to the bootstrap. New York: Chapman and Hall.

Eisner, E.W. (1983). Anastasia might still be alive, but the monarchy is dead. Educational Researcher, 12(5), 13-14, 23-34.

*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.

*Emmons, N.J., Stallings, W.M., & Layne, B.H. (1990, April). Statistical methods used in *American Educational Research Journal, Journal of Educational Psychology*, and *Sociology of Education* from 1972 through 1987. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319 797)

Ezekiel, M. (1930). Methods of correlational analysis. New York: Wiley.

Fan, X. (1996). Canonical correlation analysis as a general analytic model. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 71-94). Greenwich, CT: JAI Press.

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? Structural Equation Modeling, 4, 65-79.

Fetterman, D.M. (1982). Ethnography in educational research: The dynamics of diffusion. Educational Researcher, 11(3), 17-22, 29.

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

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. Psychological Bulletin, 70, 245-251.

Gage, N.L. (1978). The scientific basis of the art of teaching. New York: Teachers College Press.

Gage, N.L. (1985). Hard gains in the soft sciences: The case of pedagogy. Bloomington, IN: Phi Delta Kappa Center on Evaluation, Development, and Research.

Gall, M.D., Borg, W.R., & Gall, J.P. (1996). Educational research: An introduction (6th ed.). White Plains, NY: Longman.

Glantz, S.A. (1980). Biostatistics: How to detect, correct and prevent errors in the medical literature. Circulation, 61, 1-7.

Glass, G.V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

*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.

*Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1020.

Grimm, L.G., & Yarnold, P.R. (Eds.). (1995). Reading and understanding multivariate statistics. Washington, DC: American Psychological Association.

*Hall, B.W., Ward, A.W., & Comer, C.B. (1988). Published educational research: An empirical study of its quality. Journal of Educational Research, 81, 182-189.

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

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

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect sizes and related estimators. Journal of Educational Statistics, 6, 107-128.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Henard, D.H. (1998, January). Suppressor variable effects: Toward understanding an elusive data dynamic. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 416 215)

Herzberg, P.A. (1969). The parameters of cross-validation. Psychometrika Monograph Supplement, 16, 1-67.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1998). Applied statistics for the behavioral sciences (4th ed.). Boston: Houghton Mifflin.

Horst, P. (1966). Psychological measurement and prediction. Belmont, CA: Wadsworth.

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

Huberty, C.J, & Barton, R. (1989). An introduction to discriminant analysis. Measurement and Evaluation in Counseling and Development, 22, 158-168.

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

Huberty, C.J, & Pike, C.J. (in press). On some history regarding statistical testing. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5). Stamford, CT: JAI Press.

Humphreys, L.G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. Journal of Educational Psychology, 70, 873-876.

Humphreys, L.G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. Journal of Educational Psychology, 66, 464-472.

Hunter, J.E., & Schmidt, F.L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs (2nd ed.). Newbury Park, CA: SAGE.

Kaestle, C.F. (1993). The awful reputation of education research. Educational Researcher, 22(1), 23, 26-31.

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

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.

*Keselman, H.J., Huberty, C.J, Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Keselman, H.J., Kowalchuk, R.K., & Lix, L.M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. Psychometrika, 63, 145-163.

Keselman, H.J., Lix, L.M., & Kowalchuk, R.K. (1998). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.

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

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kromrey, J.D., & Hines, C.V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. Journal of Experimental Education, 64, 240-266.

Lancaster, B.P. (in press). Defining and interpreting suppressor effects: Advantages and limitations. In B. Thompson, B. (Ed.), Advances in social science methodology (Vol. 5). Stamford, CT: JAI Press.

*Lance, T., & Vacha-Haase, T. (1998, August). *The Counseling Psychologist*: Trends and usages of statistical significance testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Levin, J.R. (1998). To test or not to test H_0? Educational and Psychological Measurement, 58, 311-331.

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

Lord, F.M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample (Research Bulletin 50-110). Princeton, NJ: Educational Testing Service.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in medical research. The American Statistician, 52, 127-132.

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

Lunneborg, C.E. (1999). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Duxbury.

Manly, B.F.J. (1994). Randomization and Monte Carlo methods in biology (2nd ed.). London: Chapman and 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.

Mittag, K. (1992, January). Correcting for systematic bias in sample estimates of population variances: Why do we divide by n-1?. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 341 728)

*Mittag, K.G. (1999, April). A national survey of AERA members' perceptions of the nature and meaning of statistical significance tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Murphy, K.R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

*Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

*Nilsson, J., & Vacha-Haase, T. (1998, August). A review of statistical significance reporting in the *Journal of Counseling Psychology*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

*Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Olejnik, S.F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

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

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

*Reetz, D., & Vacha-Haase, T. (1998, August). Trends and usages of statistical significance testing in adult development and aging research: A review of *Psychology and Aging*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Rennie, K.M. (1997, January). Understanding the sampling distribution: Why we divide by n-1 to estimate the population variance. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin. (ERIC Document Reproduction Service No. ED 406 442)

Rokeach, M. (1973). The nature of human values. New York: Free Press.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1991). Meta-analytic procedures for social research (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), The handbook of research synthesis (pp. 231-244. New York: Russell Sage Foundation.

*Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal of Psychology, 55, 33-38.

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

Rossi, J.S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 176-197). Mahwah, NJ: Erlbaum.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Sanders, J.R. (1994). The process of developing national standards that meet ANSI guidelines. Journal of Experimental Education, 63, 5-12.

Saxe, G., & Schoenfeld, A. (1998). Annual meeting 1999. Educational Researcher, 27(5), 41.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

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

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

*Snyder, P.A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly, 13, 335-348.

Sprent, P. (1998). Data driven statistical methods. London: Chapman and Hall.

Tatsuoka, M.M. (1973a). An examination of the statistical properties of a multivariate measure of strength of relationship. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Tatsuoka, M.M. (1973b). Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), Review of research in education (pp. 273-319). Itasca, IL: Peacock.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park, CA: Sage.

Thompson, B. (1985). Alternate methods for analyzing data from 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. (1988a, 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. (1988b). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. Educational and Psychological Measurement, 48, 681-686.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24, 80-95.

Thompson, B. (1992a). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids. Educational and Psychological Measurement, 52, 905-911.

Thompson, B. (1992b, April). Interpreting regression results: beta weights and structure coefficients are both important. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 344 897)

Thompson, B. (1992c). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993a, April). The General Linear Model (as opposed to the classical ordinary sums of squares) approach to analysis of variance should be taught in introductory statistical methods classes. Paper presented at the annual meeting of the American Educational Research Association, Atlanta. (ERIC Document Reproduction Service No. ED 358 134)

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B. (1994c). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. In B. Thompson (Ed.), Advances in social science methodology (Vol. 3, pp. 3-27). Greenwich, CT: JAI Press.

Thompson, B. (1994d, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development (SWSRHD), Austin, TX. (ERIC Document Reproduction Service No.

ED 367 687)

Thompson, B. (1995a). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Educational and Psychological Measurement, 55, 84-94.

Thompson, B. (1995b). Review of *Applied discriminant analysis* by C.J Huberty. Educational and Psychological Measurement, 55, 340-350.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1997a). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26(5), 29-32.

Thompson, B. (1997b). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. Educational and Psychological Measurement, 57, 5-19.

Thompson, B. (1998a, April). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Document Reproduction Service No. ED 419 023) [also available on the Internet through URL: **"http://acs.tamu.edu/~bbt6147"**]

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998c). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). Educational and Psychological Measurement, 58, 332-344.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them?. Theory & Psychology, 9(2), 167-183.

*Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. Exceptional Children, 65, 329-337.

Thompson, B. (in press-a). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), Reading and understanding multivariate statistics (Vol. 2). Washington, DC: American Psychological Association.

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. Educational Psychology Review.

Thompson, B., & Borrello, G.M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

*Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. Journal of Experimental Education, 66, 75-83.

*Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. Journal of Counseling and Development, 76, 436-441.

Travers, R.M.W. (1983). How research has changed American schools: A history from 1840 to the present. Kalamazoo, MI: Mythos Press.

Tryon, W.W. (1998). The inscrutable null hypothesis. American

Psychologist, 53, 796.

Tuckman, B.W. (1990). A proposal for improving the quality of published educational research. Educational Researcher, 19(9), 22-24.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

*Vacha-Haase, T., & Ness, C. (1999). Statistical significance testing as it relates to practice: Use within *Professional Psychology*. Professional Psychology: Research and Practice, 30, 104-105.

*Vacha-Haase, T., & Nilsson, J.E. (1998). Statistical significance reporting: Current trends and usages within *MECD*. Measurement and Evaluation in Counseling and Development, 31, 46-57.

*Vockell, E.L., & Asher, W. (1974). Perceptions of document quality and use by educational decision makers and researchers. American Educational Research Journal, 11, 249-258.

Waliczek, T.M. (1996, January). A primer on partial correlation coefficients. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 393 882)

Walsh, B.D. (1996). A note on factors that attenuate the correlation coefficient and its analogs. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 21-33). Greenwich, CT: JAI Press.

Wampold, B.E., Furlong, M.J., & Atkinson, D.R. (1983). Statistical significance, power, and effect size: A response to the reexamination of reviewer bias. Journal of Counseling Psychology, 30, 459-463.

*Wandt, E. (1967). An evaluation of educational research published in journals (Report of the Committee on Evaluation of Research). Washington, DC: American Educational Research Association.

*Ward, A.W., Hall, B.W., & Schramm, C.E. (1975). Evaluation of published educational research: A national survey. American Educational Research Journal, 12, 109-128.

Wilcox, R.R. (1997). Introduction to robust estimation and hypothesis testing. San Diego: Academic Press.

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

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

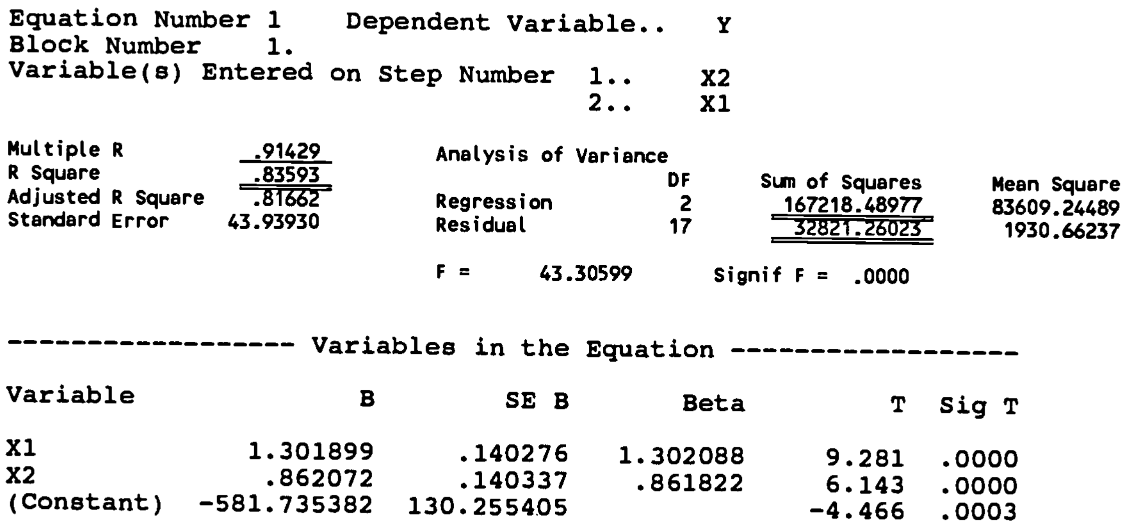
*Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4, 49-53.

Table 1
Heuristic Data Set #1 (n = 20) Involving 3 Measured Variables

ID/	Measured Variables			Synthetic/Latent Variables				
Stat.	Y	X1	X2	YHAT	yhat	yhat^2	e	e^2
1	473	392	573	422.58	-77.67	6033.22	50.42	2542.79
2	395	319	630	376.68	-123.57	15270.62	18.32	335.86
3	590	612	376	539.17	38.92	1514.44	50.83	2584.35
4	590	514	517	533.13	32.88	1081.21	56.87	3234.25
5	525	453	559	489.92	-10.33	106.65	35.08	1230.55
6	564	551	489	557.16	56.91	3239.21	6.84	46.76
7	694	722	333	645.31	145.06	21041.11	48.69	2371.37
8	356	441	531	450.16	-50.09	2508.79	-94.16	8866.10
9	408	392	531	386.37	-113.88	12968.85	21.63	467.99
10	421	551	362	447.68	-52.57	2763.51	-26.68	711.75
11	434	441	545	462.23	-38.02	1445.43	-28.23	796.87
12	342	367	489	317.61	-182.64	33355.67	24.39	594.76
13	538	465	616	554.68	54.43	2963.05	-16.68	278.28
14	369	538	390	454.89	-45.36	2057.14	-85.89	7377.44
15	499	514	489	508.99	8.74	76.45	-9.99	99.83
16	564	600	446	583.89	83.64	6995.32	-19.89	395.44
17	525	587	390	518.69	18.44	339.93	6.31	39.88
18	447	477	474	447.89	-52.36	2741.31	-0.89	0.79
19	668	648	503	695.52	195.27	38129.31	-27.52	757.08
20	603	416	757	612.44	112.19	12587.27	-9.44	89.13
Sum	10005	10000	10000	10005.00	0.00	167218.50	0.00	32821.26
M	500.25	500.00	500.00	500.25	0.00	8360.93	0.00	1641.06
SD	100.01	100.02	99.98	91.44	91.44	10739.79	40.51	2372.46

Note. These SD's are based on the population parameter formula.

Figure 1
SPSS Output of Regression Analysis
for the Table 1 Data



```
Equation Number 1    Dependent Variable..   Y
Block Number    1.
Variable(s) Entered on Step Number  1..     X2
                                    2..     X1

Multiple R            .91429       Analysis of Variance
R Square              .83593                        DF    Sum of Squares     Mean Square
Adjusted R Square     .81662       Regression        2     167218.48977      83609.24489
Standard Error      43.93930       Residual         17      32821.26023       1930.66237

                                   F =     43.30599      Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B          SE B          Beta          T   Sig T

X1             1.301899       .140276      1.302088      9.281   .0000
X2              .862072       .140337       .861822      6.143   .0000
(Constant)  -581.735382    130.255405                   -4.466   .0003
```

Note. Using an Excel function (i.e., "=FDIST(f,df1,df2)" = "=FDIST(43.30599,2,17)"), the exact $p_{CALCULATED}$ value was evaluated to be .000000213. A $p_{CALCULATED}$ value can never be 0, notwithstanding the SPSS reporting traditions for extremely small values of p; obtaining a sample with a probability of occurring of 0 would mean that you had obtained an impossible result [which is impossible to do!].

Figure 2
SPSS Output of Bivariate Product-moment Correlation Coefficients
for the 3 Measured and 2 Synthetic Variables
for Heuristic Data Set #1 (Table 1)

```
             Y         X1         X2         E         YHAT
                                                              a
Y            1.0000      .6868     -.0677      .4051      .9143
           (    20)   (    20)   (    20)   (    20)   (    20)
           P= .        P= .001    P= .777    P= .076    P= .000
                                                   b          c
X1            .6868     1.0000     -.7139      .0000      .7512
           (    20)   (    20)   (    20)   (    20)   (    20)
            P= .001    P= .       P= .000    P=1.000    P= .000
                                                   b          c
X2           -.0677     -.7139     1.0000      .0000     -.0741
           (    20)   (    20)   (    20)   (    20)   (    20)
            P= .777    P= .000    P= .       P=1.000    P= .756
                             b          b                     b
E             .4051      .0000      .0000     1.0000      .0000
           (    20)   (    20)   (    20)   (    20)   (    20)
            P= .076    P=1.000    P=1.000    P= .       P=1.000
                  a          c          c          b
YHAT          .9143      .7512     -.0741      .0000     1.0000
           (    20)   (    20)   (    20)   (    20)   (    20)
            P= .000    P= .000    P= .756    P=1.000    P= .
```

[a] The bivariate r between the Y and the $\hat{Y}$ scores is always the multiple R.

[b] The measured variables and the synthetic variable $\hat{Y}$ always have a correlation of 0 with the synthetic variable e scores.

[c] The structure coefficients for the two measured predictor variables. This can also be computed as $r_s = r$ for a given measured predictor with Y / R (Thompson & Borrello, 1985). For example, .6868 / .9143 = .7512.

Table 2
Heuristic Data Set #2 (n = 20) Involving Scores of 10 People in Each of Two Groups on 2 Measured Response Variables

Group/	Meas. Vars.		Latent	
Statistic	X	Y	Score	
1	1	0	-1.225	aeraa997.wk1 3/8/99
1	1	0	-1.225	
1	12	12	0.000	
1	12	12	0.000	
1	12	12	0.000	
1	13	11	-2.450	
1	13	11	-2.450	
1	13	11	-2.450	
1	24	23	-1.225	
1	24	23	-1.225	
2	0	1	1.225	
2	0	1	1.225	
2	11	13	2.450	
2	11	13	2.450	
2	11	13	2.450	
2	12	12	0.000	
2	12	12	0.000	
2	12	12	0.000	
2	23	24	1.225	
2	23	24	1.225	
Standardized Difference	0.137	-0.137	-2.582	
M_1	12.50	11.50	-1.23	
SD_1	7.28	7.28	0.95	
M_2	11.50	12.50	1.23	
SD_2	7.28	7.28	0.95	
M	12.00	12.00	0.00	
SD	7.30	7.30	1.55	

Note. The tabled SD values are the parameter estimates (i.e., $[SOS / \mathbf{n}]^{.5}$ = $[530.5 / 10]^{.5}$ = $53.05^{.5}$ = 7.28). The equivalent values assuming a sample estimate of the population $\underline{\sigma}$ are larger (i.e., $[SOS / \mathbf{(n-1)}]^{.5}$ = $[530.5 / 9]^{.5}$ = $58.94^{.5}$ = 7.68).

The latent variable scores were computed by applying the raw discriminant function coefficient weights, reported in Figure 3 as -1.225 and 1.225, respectively, to the two measured variables. For example, "Latent $Score_1$" or $DSCORE_1$ equals [(-1.225 x 1) +(1.225 x 0)] equals -1.225.

Figure 3
SPSS Output for 2 ANOVAs and a DDA/MANOVA
for the Table 2 Data

```
EFFECT .. GROUP
Multivariate Tests of Significance (S = 1, M = 0, N = 7 1/2)

Test Name          Value    Exact F Hypoth. DF   Error DF  Sig. of F

Pillais           .62500   14.16667       2.00      17.00       .000
Hotellings       1.66667   14.16667       2.00      17.00       .000
Wilks             .37500   14.16667       2.00      17.00       .000
Roys              .62500
Note.. F statistics are exact.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Multivariate Effect Size

TEST NAME   Effect Size

 (All)           .625

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
EFFECT .. GROUP (Cont.)
Univariate F-tests with (1,18) D. F.

Variable   Hypoth. SS   Error SS Hypoth. MS   Error MS        F  Sig. of F ETA Square

X           5.00000 1061.00000    5.00000   58.94444   .08483      .774     .00469
Y           5.00000 1061.00000    5.00000   58.94444   .08483      .774     .00469

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
EFFECT .. GROUP (Cont.)
Raw discriminant function coefficients
          Function No.

Variable             1

X               -1.225
Y                1.225

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Note. Using an Excel function (i.e., "=FDIST(f,df1,df2)" = "=FDIST(14.16667,2,17)"), the exact $p_{CALCULATED}$ value was evaluated to be .000239. A $p_{CALCULATED}$ value can never be 0, notwithstanding the SPSS reporting traditions for extremely small values of p; obtaining a sample with a probability of occurring of 0 would mean that you had obtained an impossible result [which is impossible to do!].

Figure 4
SPSS ANOVA Output for the Multivariate Synthetic Variable for the DDA/MANOVA Results

```
      Variable  DSCORE
   By Variable  GROUP
                                 Analysis of Variance

                                  Sum of        Mean                  F       F
        Source          D.F.     Squares       Squares             Ratio   Prob.

Between Groups            1       30.0125       30.0125          30.0000   .0000
Within Groups            18       18.0075        1.0004
Total                    19       48.0200
```

Note. The degrees of freedom from this ANOVA of the DDA/MANOVA synthetic variables (i.e., "DSCORE") are wrong, because the computer does not realize that the multivariate synthetic variable, "DSCORE," actually is a composite of two measured variables, and so therefore the *F* and *p* values are also wrong. However, the eta^2 can be computed as 30.0125 / 48.020 = **.625**, which exactly matches the multivariate effect size for the DDA/MANOVA reported by SPSS.

Table 3
Heuristic Data Set #3 (n = 21) Involving Scores of 21 People on One Measured Response Variable and Three Pairs of Intervally- and Nominally-Scaled Predictors

		Predictors					
Id	Y	X1	X1′	X2	X2′	X3	X3′
1	495	399	1	499	1	483	1
2	497	399	1	499	1	492	1
3	499	400	1	499	1	495	1
4	499	400	1	499	1	495	1
5	499	400	1	499	1	495	1
6	501	401	1	499	1	496	1
7	503	401	1	499	1	497	1
8	496	499	2	500	2	498	2
9	498	499	2	500	2	499	2
10	500	500	2	500	2	500	2
11	500	500	2	500	2	500	2
12	500	500	2	500	2	500	2
13	502	501	2	500	2	501	2
14	504	501	2	500	2	502	2
15	498	599	3	501	3	503	3
16	500	599	3	501	3	504	3
17	502	600	3	501	3	505	3
18	502	600	3	501	3	505	3
19	502	600	3	501	3	505	3
20	504	601	3	501	3	508	3
21	506	601	3	501	3	517	3

Note. X1′, X2′, and X3′ are the re-expressions in nominal score form of their intervally-scaled variable counterparts.

Figure 5
Regression (Y, X3) and ANOVA (Y, X3') of Table 3 Data

Regression (Y, X3)

```
Equation Number 1    Dependent Variable..   Y

Block Number     1.  Method:  Enter      X3

                              Analysis of Variance
Multiple R          .77282                          DF     Sum of Squares     Mean Square
R Square            .59725    Regression             1           91.18013        91.18013
Adjusted R Square   .57605    Residual              19           61.48654         3.23613
Standard Error     1.79893

                              F =     28.17564      Signif F =  .0000
```

ANOVA (Y, X3')

```
      Variable  Y
   By Variable  X3A
                                  Analysis of Variance

                                    Sum of         Mean           F         F
        Source            D.F.     Squares       Squares        Ratio     Prob.
Between Groups              2      32.6667       16.3333       2.4500     .1145
Within Groups              18     120.0000        6.6667
Total                      20     152.6667
```

Note. Using an Excel function (i.e., "=FDIST(f,df1,df2)" = "=FDIST(28.17564,1,19)"), the exact $p_{CALCULATED}$ value was evaluated to be .0000401. For the ANOVA, eta^2 was computed to be 21.39% (32.6667 / 152.6667).

Table 4
My Most Recent Essays Regarding Statistical Tests

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher*, *26*(5), 29-32.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, *5*(2), 33-38.

Vacha-Haase, T., & Thompson, B. (1998). Further comments on statistical significance tests. *Measurement and Evaluation in Counseling and Development*, *31*, 63-67.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*, 799-800.

Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children*, *65*, 329-337.

Thompson, B. (1999). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*(2), 193-199.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology*, *133*, 133-140.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them?. *Theory & Psychology*, *9*(2), 167-183.

Thompson, B. (in press). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Table 5
Heuristic Data Set #3 Defining a Population of $\underline{N}$=20 Scores

ID	X	
1	430	aeraa993.wk4 3/7/99
2	431	aeraa993.out
3	432	
4	433	
5	435	
6	438	
7	442	
8	446	
9	451	
10	457	
11	465	
12	474	
13	484	
14	496	
15	512	
16	530	
17	560	
18	595	
19	649	
20	840	
μ	500.00	
σ	97.73	

Table 6
The Sampling Distribution for the Mean of
n=3 Scores Drawn from the Table 5 Population of N=20 Scores

	Cases							
Sample	1	2	2	X_1	X_2	X_3	Mean	Ratio
1	1	2	3	430	431	432	431.00	-1.29
2	1	2	4	430	431	433	431.33	-1.29
3	1	2	5	430	431	435	432.00	-1.27
4	1	2	6	430	431	438	433.00	-1.26
5	1	2	7	430	431	442	434.33	-1.23
6	1	2	8	430	431	446	435.67	-1.21
7	1	2	9	430	431	451	437.33	-1.17
8	1	2	10	430	431	457	439.33	-1.14
9	1	2	11	430	431	465	442.00	-1.09
10	1	2	12	430	431	474	445.00	-1.03
11	1	2	13	430	431	484	448.33	-0.97
12	1	2	14	430	431	496	452.33	-0.89
13	1	2	15	430	431	512	457.67	-0.79
14	1	2	16	430	431	530	463.67	-0.68
15	1	2	17	430	431	560	473.67	-0.49
16	1	2	18	430	431	595	485.33	-0.27
17	1	2	19	430	431	649	503.33	0.06
18	1	2	20	430	431	840	567.00	1.26
19	1	3	4	430	432	433	431.67	-1.28
20	1	3	5	430	432	435	432.33	-1.27
21	1	3	6	430	432	438	433.33	-1.25
22	1	3	7	430	432	442	434.67	-1.22
23	1	3	8	430	432	446	436.00	-1.20
24	1	3	9	430	432	451	437.67	-1.17
25	1	3	10	430	432	457	439.67	-1.13
26	1	3	11	430	432	465	442.33	-1.08
27	1	3	12	430	432	474	445.33	-1.02
28	1	3	13	430	432	484	448.67	-0.96
29	1	3	14	430	432	496	452.67	-0.89
30	1	3	15	430	432	512	458.00	-0.79
31	1	3	16	430	432	530	464.00	-0.67
32	1	3	17	430	432	560	474.00	-0.49
33	1	3	18	430	432	595	485.67	-0.27
34	1	3	19	430	432	649	503.67	0.07
35	1	3	20	430	432	840	567.33	1.26
36	1	4	5	430	433	435	432.67	-1.26
37	1	4	6	430	433	438	433.67	-1.24
38	1	4	7	430	433	442	435.00	-1.22
39	1	4	8	430	433	446	436.33	-1.19
40	1	4	9	430	433	451	438.00	-1.16
41	1	4	10	430	433	457	440.00	-1.12
42	1	4	11	430	433	465	442.67	-1.07
43	1	4	12	430	433	474	445.67	-1.02
44	1	4	13	430	433	484	449.00	-0.96
45	1	4	14	430	433	496	453.00	-0.88

46	1	4	15	430	433	512	458.33	-0.78
47	1	4	16	430	433	530	464.33	-0.67
48	1	4	17	430	433	560	474.33	-0.48
49	1	4	18	430	433	595	486.00	-0.26
50	1	4	19	430	433	649	504.00	0.07
51	1	4	20	430	433	840	567.67	1.27
52	1	5	6	430	435	438	434.33	-1.23
53	1	5	7	430	435	442	435.67	-1.21
54	1	5	8	430	435	446	437.00	-1.18
55	1	5	9	430	435	451	438.67	-1.15
56	1	5	10	430	435	457	440.67	-1.11
57	1	5	11	430	435	465	443.33	-1.06
58	1	5	12	430	435	474	446.33	-1.01
59	1	5	13	430	435	484	449.67	-0.94
60	1	5	14	430	435	496	453.67	-0.87
61	1	5	15	430	435	512	459.00	-0.77
62	1	5	16	430	435	530	465.00	-0.66
63	1	5	17	430	435	560	475.00	-0.47
64	1	5	18	430	435	595	486.67	-0.25
65	1	5	19	430	435	649	504.67	0.09
66	1	5	20	430	435	840	568.33	1.28
67	1	6	7	430	438	442	436.67	-1.19
68	1	6	8	430	438	446	438.00	-1.16
69	1	6	9	430	438	451	439.67	-1.13
70	1	6	10	430	438	457	441.67	-1.09
71	1	6	11	430	438	465	444.33	-1.04
72	1	6	12	430	438	474	447.33	-0.99
73	1	6	13	430	438	484	450.67	-0.92
74	1	6	14	430	438	496	454.67	-0.85
75	1	6	15	430	438	512	460.00	-0.75
76	1	6	16	430	438	530	466.00	-0.64
77	1	6	17	430	438	560	476.00	-0.45
78	1	6	18	430	438	595	487.67	-0.23
79	1	6	19	430	438	649	505.67	0.11
80	1	6	20	430	438	840	569.33	1.30
81	1	7	8	430	442	446	439.33	-1.14
82	1	7	9	430	442	451	441.00	-1.11
83	1	7	10	430	442	457	443.00	-1.07
84	1	7	11	430	442	465	445.67	-1.02
85	1	7	12	430	442	474	448.67	-0.96
• • • •								
1131	16	17	18	530	560	595	561.67	1.16
1132	16	17	19	530	560	649	579.67	1.49
1133	16	17	20	530	560	840	643.33	2.69
1134	16	18	19	530	595	649	591.33	1.71
1135	16	18	20	530	595	840	655.00	2.90
1136	16	19	20	530	649	840	673.00	3.24
1137	17	18	19	560	595	649	601.33	1.90
1138	17	18	20	560	595	840	665.00	3.09
1139	17	19	20	560	649	840	683.00	3.43
1140	18	19	20	595	649	840	694.67	3.65

Figure 6
Graphic Presentation of the Sampling Distribution for the Mean of *n*=3 Scores Drawn from the Table 5 Population of *N*=20 Scores

```
Count Midpoint
 60   433 I************:**
146   446 I****************:********************
160   459 I********************:*******************
135   472 I***********************:**********
123   485 I**************************:****
 96   498 I************************   .
 97   511 I************************  .
 67   524 I*****************       .
 38   537 I**********           .
 21   550 I*****            .
 28   563 I*******      .
 50   576 I*********:***
 35   589 I******:**
 24   602 I***:**
 17   615 I**:*
 17   628 I*:**
 14   641 I:***
  6   654 I**
  4   667 I*
  1   680 I
  1   693 I
          +----+----+----+----+----+----+----+----+----+----+
          0        40        80       120       160       200
                          Histogram frequency
```

```
MEAN
Mean        500.000     Std err       1.581     Median     486.000
Mode        457.670     Std dev      53.395     Variance  2850.986
Kurtosis       .418     S E Kurt       .145     Skewness     1.052
S E Skew       .072     Range       263.670     Minimum    431.000
Maximum     694.670     Sum      570000.020

Percentile    Value     Percentile    Value     Percentile    Value
   1.00     433.330        2.00     435.000        3.00     436.670
   4.00     437.881        5.00     439.016        6.00     440.000
   7.00     441.330        8.00     442.425        9.00     443.330
  10.00     444.670       11.00     445.670       12.00     446.330
  25.00     459.000       50.00     486.000       75.00     524.502
  88.00     577.723       89.00     581.000       90.00     584.967
  91.00     588.670       92.00     592.908       93.00     597.713
  94.00     602.210       95.00     610.313       96.00     617.360
  97.00     625.440       98.00     639.729       99.00     648.865
```

Figure 7
Graphic Presentation of the Test Distribution for the Mean of n=3 Scores Drawn from the Table 5 Population of N=20 Scores

```
Count Midpoint
 39    -1.30 I********** .
143    -1.05 I***************:********************
159     -.80 I********************:*******************
147     -.55 I***********************:*************
125     -.30 I**************************:****
105     -.05 I************************** .
 99      .20 I*************************  .
 68      .45 I*****************        .
 37      .70 I*********            .
 21      .95 I*****            .
 33     1.20 I********     .
 49     1.45 I*********:**
 35     1.70 I******:**
 25     1.95 I***:**
 19     2.20 I**:**
 12     2.45 I:**
 14     2.70 I:***
  5     2.95 I*
  3     3.20 I*
  1     3.45 I
  1     3.70 I
             +----+----+----+----+----+----+----+----+----+----+
             0         40        80        120       160       200
                           Histogram frequency

TRATIO
Mean          .000     Std err      .030     Median     -.262
Mode         -.793     Std dev     1.000     Variance   1.001
Kurtosis      .418     S E Kurt     .145     Skewness   1.052
S E Skew      .072     Range       4.940     Minimum   -1.293
Maximum      3.647     Sum          .000

Percentile   Value     Percentile   Value    Percentile   Value

   1.00     -1.249        2.00     -1.218       3.00     -1.187
   4.00     -1.164        5.00     -1.143       6.00     -1.124
   7.00     -1.099        8.00     -1.079       9.00     -1.062
  10.00     -1.037       11.00     -1.018      12.00     -1.006
  25.00      -.768       50.00      -.262      75.00       .459
  88.00      1.456       89.00      1.518      90.00      1.592
  91.00      1.661       92.00      1.741      93.00      1.831
  94.00      1.915       95.00      2.067      96.00      2.199
  97.00      2.350       98.00      2.618      99.00      2.789
```

Table 7
Two Illustrative "Modern" Statistics

Id	X	X′	X-
1	430	433	--
2	431	433	--
3	432	433	--
4	433	433	433
5	435	435	435
6	438	438	438
7	442	442	442
8	446	446	446
9	451	451	451
10	457	457	457
11	465	465	465
12	474	474	474
13	484	484	484
14	496	496	496
15	512	512	512
16	530	530	530
17	560	560	560
18	595	560	--
19	649	560	--
20	840	560	--
M	500.00	480.10	473.07
Md	461.00	461.00	461.00
SD	100.27	49.34	38.98
S	2.40	0.72	1.04
K	6.54	-1.08	0.30

aera992.wk1 3/6/99

Table 8
Heuristic Data Set #4 for use in Illustrating
the Univariate Bootstrap

	Variables		
ID	Churches	Murders	Population
1	3505	1984	7322564
2	2023	1056	3485557
3	2863	921	2783726
4	1475	586	1027974
5	2011	447	1629902
6	1709	420	1585577
7	1313	373	1007618
8	1098	326	736014
9	937	211	935393
10	836	166	641432
11	997	151	589305
12	582	141	1110623
13	645	138	579396
14	663	134	983403
15	1010	132	741952
16	875	125	672971
17	615	113	723959
18	912	89	575396
19	559	88	573058
20	899	78	571059
21	329	52	567306
22	1162	49	576396
23	1372	44	643955
24	355	42	782224
25	867	37	529401
26	1129	27	574932
27	244	24	525439
28	1527	24	600499
29	909	23	569396
30	1328	22	592669
31	921	19	527432
32	982	17	602993
33	829	14	524953
34	1328	13	574039
35	1339	12	567496
36	1283	12	505955
37	1439	12	572039
38	999	11	523085
39	1052	9	568206
40	1428	7	524099
41	1345	6	526199
42	1423	6	580284

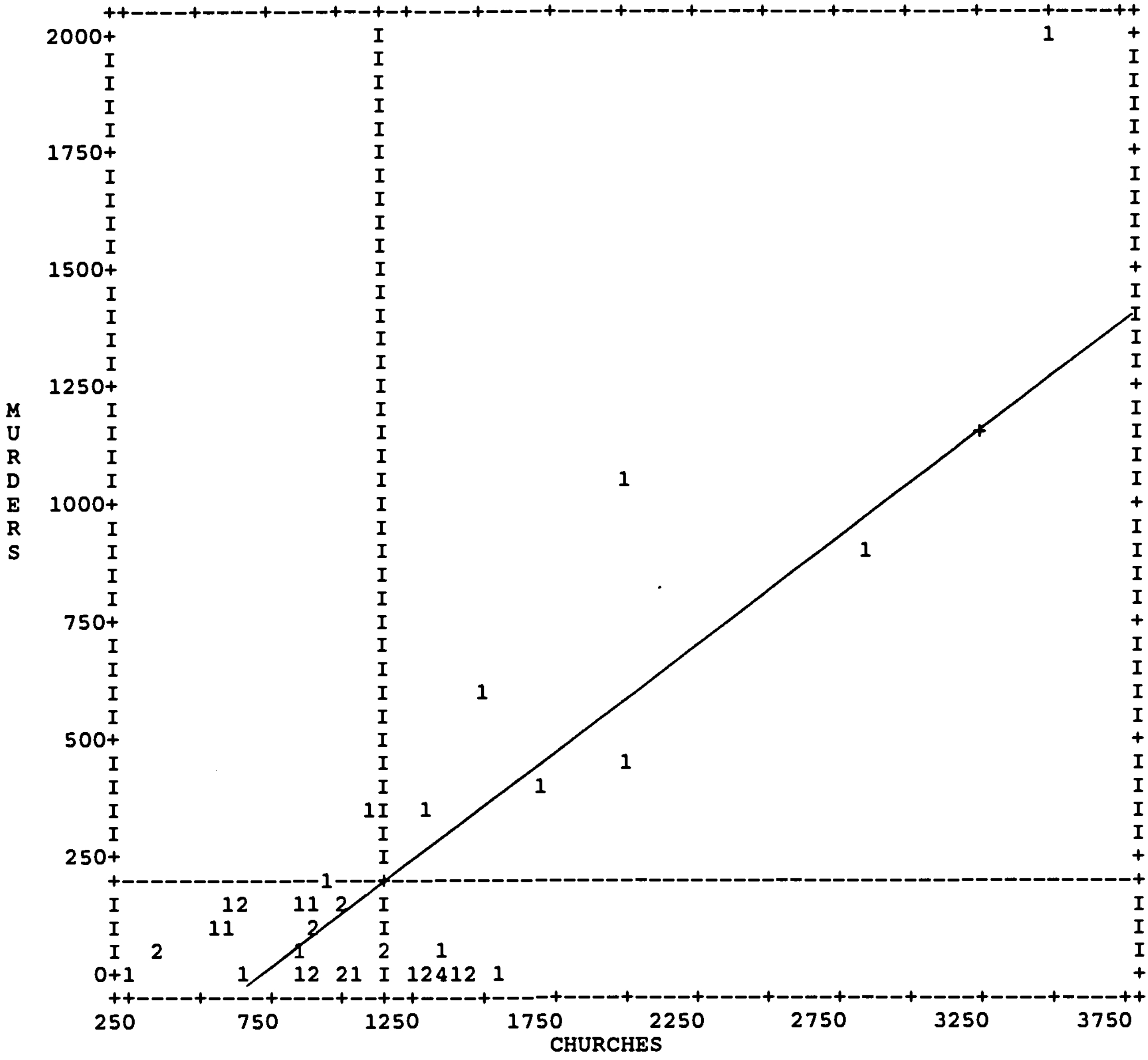
aera994.wk1 3/7/99

43	662	3	522943
44	1295	2	530299
45	1225	0	521944

Note. The first 15 cases are actual data reported by Waliczek (1996).

Figure 8
Scatterplot of the Table 8 Heuristic Data

PLOT OF MURDERS WITH CHURCHES



Note. r^2 = 60.8%; a = -362.363; b = .468.

Table 9
Some of the 1,000 Bootstrap Estimates of r

Resample	r
1	.34142710
2	.43497230
3	.59294180
4	.79517950
5	.82863380
6	.81409170
7	.82276610
8	.75451020
9	.63805250
10	.73474330
11	.71731940
12	.44586690
13	.91317640
14	.84653540
15	.86732770
• • • •	
990	.79418320
991	.57778890
992	.74192620
993	.67028270
994	.82308570
995	.78634330
996	.49483000
997	.70336210
998	.84107100
999	.77054850
1000	.76437550

Note. The actual r for the 45 pairs of scores presented in Table 8 equalled .779.

Figure 9
"Bootstrap" Estimate of the Sampling Distribution
for r with the n=45 Table 8 Data

```
r

Count Midpoint  One symbol equals approximately 8 occurrences
    0    -.500 I
    0    -.425 I
    2    -.350 I
    0    -.275 I
    0    -.200 I
    0    -.125 I
    1    -.050 I
    2     .025 I
    5     .100 I*
    5     .175 I*
   13     .250 I**
   21     .325 I:**
   27     .400 I***.
   41     .475 I*****  .
   62     .550 I********     .
   83     .625 I**********        .
  149     .700 I*******************  .
  276     .775 I********************:**************
  271     .850 I***************:******************
   42     .925 I*****    .
    0    1.000 I    .
               +----+----+----+----+----+----+----+----+
               0         80        160       240       320
                         Histogram frequency

Mean          .715     Std err       .005     Median         .768
Mode         -.384     Std dev       .166     Variance       .027
Kurtosis     5.107     S E Kurt      .155     Skewness     -1.883
S E Skew      .077     Range        1.312     Minimum       -.384
Maximum       .929     Sum        714.545

Percentile   Value     Percentile   Value     Percentile    Value

   1.00       .136        2.00       .244        3.00        .310
   4.00       .346        5.00       .367        6.00        .399
   7.00       .432        8.00       .443        9.00        .462
  10.00       .477       11.00       .495       12.00        .521
  25.00       .655       50.00       .768       75.00        .825
  88.00       .854       89.00       .856       90.00        .862
  91.00       .866       92.00       .869       93.00        .875
  94.00       .879       95.00       .883       96.00        .888
  97.00       .891       98.00       .896       99.00        .908
```

Figure 10
"Bootstrap" Estimate of the Sampling Distribution
for r-to-Z with the n=45 Table 8 Data

```
r_TO_Z

Count Midpoint  One symbol equals approximately 4 occurrences
    1      -.4 I
    1      -.3 I
    0      -.2 I
    0      -.1 I
    2       .0 I*
    6       .1 I:*
   10       .2 I:**
   17       .3 I**:*
   29       .4 I*****:*
   47       .5 I*********:**
   48       .6 I************   .
   60       .7 I***************       .
   82       .8 I*********************       .
  114       .9 I*****************************  .
  145      1.0 I********************************:***
  150      1.1 I*****************************:********
  150      1.2 I***********************:**************
   67      1.3 I*****************.
   51      1.4 I***********:*
   15      1.5 I****  .
    5      1.6 I*  .
               +----+----+----+----+----+----+----+----+----+
                 0        40        80       120       160
                         Histogram frequency

Mean          .961     Std err      .010     Median      1.016
Mode         -.405     Std dev      .302     Variance     .091
Kurtosis      .591     S E Kurt     .155     Skewness    -.733
S E Skew      .077     Range       2.052     Minimum     -.405
Maximum      1.648     Sum       960.752

Percentile   Value     Percentile  Value     Percentile  Value
   1.00       .137        2.00      .249        3.00      .321
   4.00       .361        5.00      .385        6.00      .422
   7.00       .463        8.00      .475        9.00      .499
  10.00       .519       11.00      .542       12.00      .578
  25.00       .784       50.00     1.016       75.00     1.171
  88.00      1.271       89.00     1.278       90.00     1.301
  91.00      1.317       92.00     1.329       93.00     1.356
  94.00      1.370       95.00     1.391       96.00     1.412
  97.00      1.428       98.00     1.453       99.00     1.517
```

Table 10
Percentiles of Resampled Sampling Distributions for
r-to-Z Values for the Table 8 Data with
100 and 1,000 Resamples

aeraa996.wk1 3/8/99

%ile/	n of Resamples		
Statistic	100	1000	Difference
99	1.453	1.517	-0.064
98	1.424	1.453	-0.029
97	1.405	1.428	-0.023
96	1.393	1.412	-0.019
95	1.352	1.391	-0.039
94	1.341	1.370	-0.029
93	1.309	1.356	-0.047
92	1.298	1.329	-0.031
91	1.294	1.317	-0.023
90	1.289	1.301	-0.012
89	1.284	1.278	0.006
88	1.279	1.271	0.008
75	1.195	1.171	0.024
Mean	1.000	0.961	0.039
(SD)	(0.267)	(0.302)	
50	1.063	1.016	0.047
25	0.783	0.784	-0.001
12	0.657	0.578	0.079
11	0.654	0.542	0.112
10	0.634	0.519	0.115
9	0.567	0.499	0.068
8	0.553	0.475	0.078
7	0.543	0.463	0.080
6	0.494	0.422	0.072
5	0.476	0.385	0.091
4	0.416	0.361	0.055
3	0.362	0.321	0.041
2	0.337	0.249	0.088
1	0.303	0.137	0.166

Figure 11
DDA/MANOVA Results for Sir Ronald Fisher's (1936) Iris Data
(_k_ groups = 3; _n_ = 150; _p_ response variables = 4)

DISCRIMINANT FUNCTION CENTROIDS

	Function	
Group	I	II
1	-5.50285	6.87673
2	3.93011	5.93367
3	7.88771	7.17438

DISCRIMINANT FUNCTION COEFFICIENTS

Response	Function I Coefs		Function II Coefs	
Variables	Function	r_s	Function	r_s
X1	-0.82940	0.11458	0.02410	0.16000
X2	-1.53458	-0.04043	2.16458	0.29337
X3	2.20126	0.30383	-0.93196	0.07217
X4	2.81058	0.12958	2.83931	0.15087

Figure 12
Bootstrap Resampling of Cases for the First Resample
(k groups = 3; n = 150; p response variables = 4)

RESAMPLING #1

Resample			Variable			
Step	Case	Group	X1	X2	X3	X4
1	22	1.0	5.1	3.7	1.5	0.4
2	15	1.0	5.8	4.0	1.2	0.2
3	12	1.0	4.8	3.4	1.6	0.2
4	40	1.0	5.1	3.4	1.5	0.2
5	**27**	**1.0**	**5.0**	**3.4**	**1.6**	**0.4**
6	6	1.0	5.4	3.9	1.7	0.4
7	19	1.0	5.7	3.8	1.7	0.3
8	**27**	**1.0**	**5.0**	**3.4**	**1.6**	**0.4**
9	42	1.0	4.5	2.3	1.3	0.3
10	35	1.0	4.9	3.1	1.5	0.2
***						
/abridged						
*************************************						
148	114	3.0	5.7	2.5	5.0	2.0
149	107	3.0	4.9	2.5	4.5	1.7
150	103	3.0	7.1	3.0	5.9	2.1

Figure 13
Resampling Estimates of Statistics for Resamples #1 and #2000
(*k* groups = 3; *n* = 150; *p* response variables = 4)

Resample #1

```
FUNCTION MATRIX BEFORE ROTATION
    1     -1.53727   0.40113
    2     -0.86791   1.88272
    3      3.02954  -0.96330
    4      1.89711   2.60662

FUNCTION MATRIX AFTER ROTATION
    1     -1.51555   0.47666
    2     -0.77374   1.92334
    3      2.97819  -1.11194
    4      2.02369   2.50962

STRUCTURE MATRIX BASED ON ROTATED FUNCTION
    1      0.12341   0.24790
    2     -0.02461   0.34791
    3      0.30382   0.12217
    4      0.13003   0.13888
```

Resample #2000

```
FUNCTION MATRIX BEFORE ROTATION
    1     -1.04205  -0.14641
    2     -1.22630   1.76173
    3      2.47935  -1.28000
    4      2.63734   3.88797

FUNCTION MATRIX AFTER ROTATION
    1     -1.05126  -0.04647
    2     -1.05290   1.87054
    3      2.34614  -1.51038
    4      2.99575   3.61905

STRUCTURE MATRIX BASED ON ROTATED FUNCTION
    1      0.10285   0.07936
    2     -0.02377   0.27628
    3      0.28983  -0.00799
    4      0.13456   0.13121
```

Figure 14
Map of Participant Selection Across 2,000 Resamples
(*k* groups = 3; *n* = 150; *p* response variables = 4)

Participant	Times
1	2033
2	1990
3	2031
4	1991
5	2003
6	1985
7	2001
8	2026
9	2103
10	1958
11	1968
12	2041
. . . .	
146	1942
147	1958
148	1981
149	1906
150	2036
Min	1892
Max	2125
Mean	1999.99
SD	44.56

Figure 15
Mean (and SD) of Bootstrap Estimates Across 2,000 Resamples
(k groups = 3; n = 150; p response variables = 4)

*** SUMMARY STATISTICS FOR GROUP CENTROIDS:

Group	Statistic	Function	
		I	II
1	M	-5.65456	6.79552
	SD	1.16636	1.86960
	S	1.63016	-0.26639
	K	17.38350	-0.05702
2	M	4.01345	5.81133
	SD	1.13114	1.87885
	S	-1.00750	-0.25797
	K	9.06839	-0.05715
3	M	8.06143	7.10633
	SD	1.38417	1.86440
	S	-3.82083	-0.26536
	K	49.27710	-0.05739

Var	Statistic	Function I		Function II	
		Function	r_s	Function	r_s
X1	M	-0.84924	0.10844	0.00669	0.15045
	SD	0.29921	0.01392	0.61489	0.08020
X2	M	-1.60807	-0.04132	2.15484	0.27868
	SD	0.39638	0.02294	0.46866	0.02998
X3	M	2.24008	0.29401	-0.93826	0.06939
	SD	0.29062	0.01910	0.66124	0.06334
X4	M	2.91339	0.12619	2.86778	0.14410
	SD	0.40405	0.01246	0.71420	0.01408

Table 11
Effect Size Reporting Practices Described in 11 Empirical Studies of the Quantitative Studies Published in 23 Journals

Empirical Study	Journals Studied	Years	Effects Reported
1. Keselman et al. (1998)			
	Between-subjects Univariate	1994-1995	9.8%
	American Education Research Journal		
	Child Development		
	Cognition and Instruction		
	Contemporary Educational Psychology		
	Developmental Psychology		
	Educational Technology, Research and Development		
	Journal of Counseling Psychology		
	Journal of Educational Computing Technology		
	Journal of Educational Psychology		
	Journal of Experimental Child Psychology		
	Sociology of Education		
	Between-subjects Multivariate	1994-1995	10.1%
	American Education Research Journal		
	Child Development		
	Developmental Psychology		
	Journal of Applied Psychology		
	Journal of Counseling Psychology		
	Journal of Educational Psychology		
2. Kirk (1996)			
	Journal of Applied Psychology	1995	23.0%
	Journal of Educational Psychology	1995	45.0%
	Journal of Experimental Psychology	1995	88.0%
	Journal of Personality and Social Psychology	1995	53.0%
3. Lance & Vacha-Haase (1998)	The Counseling Psychologist	1995-1996	40.5%
4. Nilsson & Vacha-Haase (1998)	Journal of Counseling Psychology	1995-1997	53.2%
5. Reetz & Vacha-Haase (1998)	Psychology and Aging	1995-1997	46.9%
6. Snyder & Thompson (1998)	School Psychology Quarterly	1990-1996	54.3%
7. Thompson (1999b)	Exceptional Children	1996-1998	13.0%
8. Thompson & Snyder (1997)	Journal of Experimental Education	1994-1997	36.4%
9. Thompson & Snyder (1998)	Journal of Counseling and Development	1996	10.0%
10. Vacha-Haase & Ness (1999)	Professional Psychology: Research and Practice	1995-1997	21.2%
11. Vacha-Haase & Nilsson (1998)	Measurement & Evaluation in Counseling and Development	1990-1996	35.3%

Table 12
Heuristic Literature #1

k	n	p_{calc}	F_{calc}	$omega^2$	eta^2	SOS_{exp}	df_{ex}	MS_{exp}	SOS_{unex}	df_{un}	MS_{unexp}
1	40	.0479	4.18	7.4%	9.9%	5.5	1	5.50	50	38	1.32
2	30	.5668	.34	-2.3%	1.2%	.6	1	.60	50	28	1.79
3	40	.4404	.61	-1.0%	1.6%	.8	1	.80	50	38	1.32
4	50	.3321	.96	-0.1%	2.0%	1.0	1	1.00	50	48	1.04
5	60	.2429	1.39	.6%	2.3%	1.2	1	1.20	50	58	.86
6	30	.2467	1.40	1.3%	4.8%	2.5	1	2.50	50	28	1.79
7	40	.1761	1.90	2.2%	4.8%	2.5	1	2.50	50	38	1.32
8	50	.1279	2.40	2.7%	4.8%	2.5	1	2.50	50	48	1.04
9	60	.0939	2.90	3.1%	4.8%	2.5	1	2.50	50	58	.86
10	70	.0696	3.40	3.3%	4.8%	2.5	1	2.50	50	68	.74
11	9	0.9423	0.06	-26.4%	2.0%	2.0	2	1.00	100	6	16.67
12	12	0.9147	0.09	-17.9%	2.0%	2.0	2	1.00	100	9	11.11
13	15	0.8880	0.12	-13.3%	2.0%	2.0	2	1.00	100	12	8.33
14	18	0.8620	0.15	-10.4%	2.0%	2.0	2	1.00	100	15	6.67
15	21	0.8368	0.18	-8.5%	2.0%	2.0	2	1.00	100	18	5.56
16	24	0.8123	0.21	-7.0%	2.0%	2.0	2	1.00	100	21	4.76
17	27	0.7885	0.24	-6.0%	2.0%	2.0	2	1.00	100	24	4.17
18	30	0.7654	0.27	-5.1%	2.0%	2.0	2	1.00	100	27	3.70
19	33	0.7430	0.30	-4.4%	2.0%	2.0	2	1.00	100	30	3.33
20	36	0.7213	0.33	-3.9%	2.0%	2.0	2	1.00	100	33	3.03
Min		0.0479		-26.4%	1.2%						
Max		0.9423		7.4%	9.9%						
M		0.5309		-4.3%	3.0%						
SD		0.3215		7.9%	2.0%						

aera9922.wk1 3/12/99

Table 13
Heuristic Literature #2

k	n	p_{calc}	F_{calc}	$omega^2$	eta^2	SOS_{exp}	df_{ex}	MS_{exp}	SOS_{unex}	df_{un}	MS_{unexp}
1	6	.0601	6.75	48.9%	62.8%	84.4	1	84.40	50	4	12.50
2	8	.0600	5.35	35.2%	47.1%	44.6	1	44.60	50	6	8.33
3	10	.0602	4.78	27.5%	37.4%	29.9	1	29.90	50	8	6.25
4	12	.0599	4.50	22.6%	31.0%	22.5	1	22.50	50	10	5.00
5	14	.0598	4.32	19.2%	26.5%	18.0	1	18.00	50	12	4.17
6	16	.0604	4.17	16.5%	23.0%	14.9	1	14.90	50	14	3.57
7	18	.0600	4.10	14.7%	20.4%	12.8	1	12.80	50	16	3.13
8	20	.0599	4.03	13.2%	18.3%	11.2	1	11.20	50	18	2.78
9	22	.0604	3.96	11.9%	16.5%	9.9	1	9.90	50	20	2.50
10	24	.0605	3.92	10.8%	15.1%	8.9	1	8.90	50	22	2.27
11	9	.0603	4.65	44.8%	60.8%	155.0	2	77.50	100	6	16.67
12	12	.0601	3.91	32.6%	46.5%	86.8	2	43.40	100	9	11.11
13	15	.0601	3.59	25.7%	37.4%	59.8	2	29.90	100	12	8.33
14	18	.0601	3.41	21.1%	31.3%	45.5	2	22.75	100	15	6.67
15	21	.0600	3.30	18.0%	26.8%	36.7	2	18.35	100	18	5.56
16	24	.0601	3.22	15.6%	23.5%	30.7	2	15.35	100	21	4.76
17	27	.0601	3.17	13.8%	20.9%	26.4	2	13.20	100	24	4.17
18	30	.0605	3.12	12.4%	18.8%	23.1	2	11.55	100	27	3.70
19	33	.0602	3.09	11.2%	17.1%	20.6	2	10.30	100	30	3.33
20	36	.0599	3.07	10.3%	15.7%	18.6	2	9.30	100	33	3.03
Min		.0598		10.3%	15.1%						
Max		.0605		48.9%	62.8%						
<u>M</u>		.0601		21.3%	29.8%						
<u>SD</u>		.0002		11.0%	14.2%						

aera9921.wk1 3/12/99

Table 14
"Shrinkage" as a Function of $\underline{n}$, $\underline{n}_{predictors}$, $\underline{R}^2$

R^2=50%; n_{pv}=3		R^2=50%; n=50		n=50; n_{pv}=3	
n	R^2*	n_{pv}	R^2*	R^2	R^2*
5	-100.00%	45	-512.50%	0.01%	-6.51%
7	0.00%	35	-75.00%	0.10%	-6.42%
10	25.00%	25	-2.08%	1.00%	-5.46%
15	36.36%	15	27.94%	5.00%	-1.20%
20	40.63%	10	37.18%	10.00%	4.13%
25	42.86%	9	38.75%	15.00%	9.46%
30	44.23%	8	40.24%	20.00%	14.78%
45	46.34%	7	41.67%	25.00%	20.11%
50	46.74%	6	43.02%	30.00%	25.43%
75	47.89%	5	44.32%	35.00%	30.76%
100	48.44%	4	45.56%	50.00%	46.74%
500	49.70%	3	46.74%	75.00%	73.37%
1000	49.85%	2	47.87%	90.00%	89.35%
10000	49.98%	1	48.96%	99.00%	98.93%

aera9930.wk1 3/13/99

Note. $\underline{R}^2$* = "adjusted $\underline{R}^2$.

Table 15
"Shrinkage" as an Interaction Effect

Combination				
n	n_{pv}	R^2	R^2*	Shrinkage
56	3	13.8%	8.8%	5.0%
93	5	13.8%	8.8%	5.0%
128	7	13.8%	8.8%	5.0%
166	9	13.8%	8.8%	5.0%
200	11	13.8%	8.8%	5.0%
19	1	13.8%	8.7%	5.1%
38	2	13.8%	8.9%	4.9%
56	3	13.8%	8.8%	5.0%
74	4	13.8%	8.8%	5.0%
93	5	13.8%	8.8%	5.0%
50	4	44.0%	39.0%	5.0%
40	6	72.5%	67.5%	5.0%
30	8	87.0%	82.0%	5.0%

aera9931.wk1 3/13/99

Note. *R^2** = "adjusted *R^2*. The 13.8% effect size is the value that Cohen (1988, pp. 22-27) characterized as "large," at least as regards result typicality.

Table 16
Selected Features of "Classical" and "Modern" Inquiry

"Classical" Research Model	Problem	"Modern" Research Model
Introduction Prior literature is characterized only as regards whether or not previous results were statistically significant.	Because p values are confounded indices jointly influenced by n's and effect sizes and other factors, conclusions based on "vote counts" integrate apples-and-oranges comparisons.	Characterizations of previous literature include summaries of prior effect sizes (potentially including "corrected" effect indices that adjust for estimated sampling error).
Hypothesis Formulation Researchers employ a mindless "point and click" specification of "nil" null hypotheses.	Hypotheses are not formulated based on a reflective integration of theory and specific previous results.	Null hypotheses are grounded in specific results from previous studies.
Discussion Results are only evaluated for statistical significance.	Studies with large sample sizes and small effects are overinterpreted; studies with small sample sizes and large effects are underinterpreted (cf. Wampold, Furlong & Atkinson, 1983).	Effect sizes are always reported and interpreted in relation to (a) the valuing of the outcome variable(s) and (b) the effects reported in the prior related literature.

Appendix A
SPSS for Windows Syntax Used to Analyze the Table 1 Data

```
SET blanks=-99999 printback=listing .
TITLE 'AERAA991.SPS     **********************************'.
DATA LIST
  FILE='c:\aeraad99\aeraa991.dta' FIXED RECORDS=1 TABLE /1
  ID 1-2 Y 9-11 X1 18-20 X2 27-29 .
list variables=all/cases=9999 .

descriptives
  variables=all/statistics=mean stddev skewness kurtosis .
correlations
  variables=Y X1 X2/statistics=descriptives .
regression variables=y x1 x2/dependent=y/
  enter x1 x2 .

subtitle '1  show synthetic vars are the focus of all analyses'.
compute yhat= -581.735382 +(1.301899 * x1) +(.862072 * x2) .
compute e=y-yhat .
print formats yhat e (F8.2) .
list variables=all/cases=9999 .
correlations variables=y x1 x2 e yhat/
  statistics=descriptives .
```

Appendix B
SPSS for Windows Syntax Used to Analyze the Table 2 Data

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=listing.
TITLE 'AERAA997.SPS   ANOVA/MANOA ###############' .
DATA LIST
  FILE='c:\spsswin\aeraa997.dta' FIXED RECORDS=1 TABLE/1
  group 12 x 18-19 y 25-26 .
list variables=all/cases=99999/format=numbered .

oneway x y by group(1,2)/statistics=all .
manova x y by group(1,2)/
  print signif(mult univ) signif(efsize) cellinfo(cov)
  homogeneity(boxm)/discrim raw stan cor alpha(.99)/
  design .

compute dscore=(-1.225 * x) + (1.225 * y) .
print formats dscore(F8.3) .
list variables=all/cases=99999/format=numbered .
oneway dscore by group(1,2)/statistics=all .
```

Appendix C
SPSS for Windows Syntax Used to Analyze the Table 3 Data

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=listing.
TITLE 'AERA9910.SPS   Var Discard ###################' .
DATA LIST
  FILE='a:aera9910.dta' FIXED RECORDS=1 TABLE/1
  id 1-2 y 7-9 x3 14-16 x3a 20 .
list variables=all/cases=99999/format=numbered .

descriptives variables=all/statistics=all .
regression variables=y x3/dependent=y/enter x3 .
oneway y by x3a(1,3)/statistics=all .
```

AERA
* INVITED ADDRESS*

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

ERIC®

TM029652

I. DOCUMENT IDENTIFICATION:

Title: COMMON METHODOLOGY MISTAKES IN EDUCATIONAL RESEARCH, REVISITED. ALONG WITH A PRIMER ON BOTH EFFECT SIZES AND THE BOOTSTRAP

Author(s): BRUCE THOMPSON

Corporate Source:	Publication Date: 4/22/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRUCE THOMPSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ☐

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Bruce	Position: PROFESSOR
Printed Name: BRUCE THOMPSON	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1335
	Date: 3/30/99

OVER