DOCUMENT RESUME

ED 325 509 TM 015 747

AUTHOR Thompson, Bruce; Melancon, Janet G.
TITLE Bootstrap versus Statistical Effect Size Corrections: A Comparison with Data from the Finding Embedded Figures Test.
PUB DATE 1 Nov 90
NOTE 30p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (19th, New Orleans, LA, November 14-16, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Assisted Testing; *Correlation; *Effect Size; Error of Measurement; Estimation (Mathematics); Higher Education; Meta Analysis; Research Methodology; *Sampling; *Statistical Analysis; *Test Interpretation; Undergraduate Students
IDENTIFIERS *Bootstrap Methods; *Finding Embedded Figures Test; Group Embedded Figures Test

ABSTRACT

Effect sizes have been increasingly emphasized in research as more researchers have recognized that: (1) all parametric analyses (t-tests, analyses of variance, etc.) are correlational; (2) effect sizes have played an important role in meta-analytic work; and (3) statistical significance testing is limited in its capacity to inform scientific inquiry. However, effect sizes tend to be biased by sampling and measurement error. The performance of the statistical corrections for sampling error bias of R. J. Wherry and P. A. Herzberg is illustrated and reviewed. The corrections are compared with empirical estimates of sampling error derived using "bootstrap" methods. A data set involving the responses of 31 college undergraduates (18 females and 13 males) on the Finding Embedded Figures Test (FEFT) and the Group Embedded Figures Test, is used for illusrative purposes to make the discussion concrete. It is suggested that bootstrap methods provide important insights for the researcher and are readily accessible to researchers due to the availability of user-friendly computer programs that automate the procedure (i.e., programs designed for use on microcomputers). Seven tables illustrate the example. An appendix provides an item analysis for heuristic FEFT data. A list of 65 references is included. (Author/SLD)

boostrap.wp0 11/1/90

BOOTSTRAP VERSUS STATISTICAL EFFECT SIZE CORRECTIONS: A COMPARISON WITH DATA FROM THE FINDING EMBEDDED FIGURES TEST

Bruce Thompson

Texas A&M University 77843-4225

Janet G. Melancon

Loyola University

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, November 14, 1990.

ABSTRACT

Effect sizes have been increasingly emphasized in research, as more researchers have recognized: (a) that all parametric analyses (t-tests, ANOVA, etc.) are correlational, (b) that effect sizes have played an important role in meta-analytic work, and (c) that statistical significance testing is limited in its capacity to inform scientific inquiry. But effect sizes tend to be biased by sampling and measurement error. The paper illustrates and reviews the performance of the Wherry and the Herzberg statistical corrections for sampling error bias, and compares the corrections with empirical estimates of sampling error derived using "bootstrap" methods. A data set involving responses of 31 subjects on the Finding Embedded Figures Test and the Group Embedded Figures Test is used for illustrative purposes, to make the discussion concrete. It is suggested that "bootstrap" methods provide the researcher with important insights, and are today readily accessible to researchers thanks to the availability of user-friendly computer programs that automate the procedure, some of which have been written for popular microcomputers.

Researchers have increasingly emphasized the examination of effect sizes as a focal part of interpreting empirical results. Many effect size estimates (e.g., Hays, 1981; Tatsuoka, 1973) are available for researchers who wish to garner some insight regarding result importance. The simplest effect sizes are analogous to the coefficient of determination (r^2). For example, in analysis of variance the sum of squares (SOS) for an effect can be divided by the SOS total to compute the correlation ratio (also called eta squared), just as the SOS explained in regression divided by the SOS total is the squared multiple correlation coefficient. Such statistics inform the researcher regarding what proportion of variance in the dependent variable(s) is explained by a given predictor. The simplest effect sizes are based on the data in hand and sample size is not considered as part of the calculations.

Three factors have led to the increased emphasis on effect size interpretation. First, researchers have increasingly recognized that all parametric analytic methods (t-tests, ANOVA, ANCOVA, MANOVA, etc.) are correlational, i.e., are special cases of canonical correlation analysis (Knapp, 1978; Thompson, 1984). Thus, canonical correlation analysis can be used to implement all parametric analyses, as Thompson (1988a) illustrates, just as all univariate parametric methods can be implemented as regression analyses (Cohen, 1968; Thompson, 1985). This recognition has stimulated researchers to realize that effect sizes analogous to squared correlation coefficients are just as important to interpret with experimental designs as squared correlation coefficients are

to interpret with correlational designs.

Second, researchers have increasingly recognized the legitimacy and the utility of meta-analytic methods. Though popularized by Glass and his colleagues (Glass, 1976; Glass, McGaw & Smith, 1981), meta-analytic methods actually date back to work by Fisher (1932), by Cochran (1937, 1943), and especially to work by Rosenthal (1963, 1984). Kulik and Kulik (in press) provide an excellent review of this history. One effect of popularized meta-analytic methods has been the popularization of effect size computations like those derived in meta-analysis.

Third, researchers have increasingly recognized that statistical significance testing is extremely limited in its capacity to inform scientific inquiry (Carver, 1978; Chow, 1988; Huberty, 1987; Kupfersmid, 1988; Rosnow & Rosenthal, 1989; Thompson, 1988c, 1989a, 1989b). Even some widely respected authors of prominent textbooks are sometimes not quite sure what role significance tests should play in analysis (Thompson, 1987a, 1988e), and some dissertation authors too may be disproportionately susceptible to excessive awe for significance tests (Eason & Daniel, 1989; Thompson, 1988b). Researchers who have had the fortunate experience of working with large samples (cf. Kaiser, 1976) soon realize that virtually all null hypotheses will be rejected, since "the null hypothesis of no difference is almost never exactly true in the population" (Thompson, 1987b, p. 14). As Meehl (1978, p. 822) notes, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the

null hypothesis, taken literally, is always false." Thus Hays (1981, p. 293) argues that "virtually any study can be made to show significant results if one uses enough subjects." The recognition that statistical significance testing is largely a test of sample size, which size the researcher already knew prior to conducting a significance test, has led to an increased emphasis on effect size interpretation.

Because the simpler effect sizes (e.g., eta squared) capitalize on sampling error as part of their inherent least-squares or correlational logic, the simpler effect sizes do overestimate both the effect size in the full population and the effect size likely to be realized in future studies. But correction formulas (Maxwell, Camp & Arvey, 1981; Rosnow & Rosenthal, 1988) can be applied to estimate population effect sizes based on sample results (e.g., Wherry, 1931), or to estimate the effect size estimates likely in future samples (Herzberg, 1969). Correction formulas are also available to adjust for attenuation due to measurement features such as limited reliability of measurement (Guilford, 1954, p. 400) or restricted variability of measurement (Borg & Gall, 1989, pp. 598-599).

The purpose of the present paper is to review two effect size estimates (Herzberg, 1969; Wherry, 1931) that can be computed to adjust for sample size influences (called "shrinkage" corrections, since the corrected effect size estimates tend to be smaller in size than uncorrected estimates), and to compare these two theoretically derived statistical corrections requiring assumptions

about the form of sampling error influences with methods that make fewer assumptions and instead ground estimates in a more thorough empirical examination of the data in hand. With respect to estimates in the second genre, the present study will review the "bootstrap" logics developed by Efron and his colleagues. A small data set (n=31) is employed for heuristic purposes to make the discussion more concrete.

Heuristic Example

Data from 31 subjects who completed both the Finding Embedded Figures Test (FEFT) (Melancon & Thompson, 1987, 1989, 1990a, 1990b, 1990c; Thompson & Melancon, 1990) and the widely known Group Embedded Figures Test (GEFT) are employed here for illustrative purposes. These data have not been previously reported.

The 31 (18 females; 58.15%) subjects were undergraduate college students. The mean age of the subjects was 20.7 (SD=4.7). The means for number of correct answers on the four variables of interest in the present example were: (a) number of right answers on the 35 items in FEFT Part A, 27.9 (SD=4.9); (b) number of right answers on the 35 items in FEFT Part B, 26.8 (SD=4.7); (c) number of right answers on the full 70 items of the FEFT, 54.7 (SD=9.0); and (d) number of right answers on the 18 items of the GEFT, 11.4 (SD=5.1).

The six unique correlation coefficients ((v*(v-1))/2 = (4*3)/2) among the four variables were of interest from a measurement point of view. The correlation between FEFT total scores and GEFT total scores was of particular interest, since the

result is a concurrent validity coefficient. The calculated correlation matrix is presented in Table 1.

INSERT TABLE 1 ABOUT HERE.

Corrections for Shrinkage in Estimating Population Effect Size Using Sample Results

Various correction formulas are available to adjust for expected "shrinkage" when estimating population effect size using sample results (e.g., Olkin & Pratt, 1958). However, Carter (1979) notes that the various corrections tend to yield very similar results, especially when sample sizes are greater than 50. The Wherry (1931) correction formula is probably the most widely used, e.g., this is the correction SPSS-X uses to compute the "adjusted" squared multiple correlation coefficient. Given v predictor variables (v=1 in the bivariate case involving a single predictor) and n subjects, the Wherry correction can be expressed as:

$$R^2 - ((1 - R^2) * (v / (n - v - 1))),$$

or equivalently as:

$$1 - ((n - 1) / (n - v - 1)) * (1 - R^2).$$

Thus, for the correlation (r=.5120) between FEFT total scores and GEFT total scores for the 31 subjects, the correction would be:

```
r      **2 - ((1 - r      **2) * ( v / ( n - v -1)))
.5120  **2 - ((1 - .5120  **2) * ( 1 / (31 - 1 -1)))
.262144   - ((1 - .262144  ) * ( 1 / (    29    )))
.262144   - ((    .737856  ) * (     .034482     ))
.262144   - (            0.025443                 )
.236700
```

Table 2 presents all six unique bivariate correlation coefficients for the 31 subjects, after adjustment for "shrinkage" using the

Wherry algorithm.

INSERT TABLE 2 ABOUT HERE.

Corrections for Shrinkage in Estimating Replication Effect Size Expected in Future Studies, Based on Sample Results

Stevens (1986, pp. 78-84) incisively implies that researchers usually ground their work in empirical findings from previous samples, and in actual practice usually want their work to generalize to future samples in future research rather than to the unknowable population. Herzberg (1969) provides a correction for this estimate that also might be used in creating coefficient aggregates to evaluate variable importance:

$$1 - ((n-1)/(n-v-1))((n-2)/(n-v-2))((n+1)/n)(1-R^2).$$

Thus, for the correlation (r=.5120) between FEFT total scores and GEFT total scores for the 31 subjects, the correction would be:

```
1 - (( n-1)/( n-v-1))*(( n-2)/( n-v-2))*(( n+1)/ n)*(1- r    **2)
1 - ((31-1)/(31-1-1))*((31-2)/(31-1-2))*((31+1)/31)*(1- .5120**2)
1 - (( 30 )/(  29  ))*(( 29 )/(  28  ))*(( 32 )/31)*(1- .262144 )
1 - (   1.034482    )*(   1.035714    )*( 1.032258)*(   .737856 )
1 - (                  1.071428    )*( 1.032258)*(   .737856 )
1 -                                    ( 1.105990 *   .737856 )
1 -                                                   .816061
                                                      .183938
```

Table 3 presents all six unique bivariate correlation coefficients for the 31 subjects, after adjustment for "shrinkage" using the Herzberg algorithm.

INSERT TABLE 3 ABOUT HERE.

"Bootstrap" Empirical Estimates of Effect Size Stability

A third strategy, like the Herzberg correction, emphasizes interpretation based on estimated likelihood that results will replicate. This emphasis is compatible with the basic purpose of science: isolating conclusions that replicate under stated conditions. Notwithstanding common misconceptions to the contrary, significance tests do not evaluate the probability that results will generalize (Carver, 1978; Thompson, 1987b).

The "bootstrap" methods developed by Efron and his colleagues (cf. Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1987, in press) are extremely powerful. Most conventional statistical estimates invoke the concept of the standard error (SE) of the statistic of interest, i.e., the standard deviation of the error of the estimates of population parameters. Typically, an assumption is made that standard errors are randomly and normally distributed during the sampling process, and the SE is derived statistically rather than empirically.

For example, for all bivariate correlation coefficients (r) expressed after the Fisher r-to-Z transformation and involving a sample size of 31, the standard error (Glass & Hopkins, 1984, p. 305) is taken to be

$$1 / (n - 3) ** .5.$$

For the correlation (r=.5120) between FEFT total scores and GEFT total scores for the 31 subjects, the correlation expressed as Z would be:

$$
\begin{aligned}
|Z| &= .5 \ln ((1 + |\ r\ |) / (1 - |\ r\ |)) \\
&0.5 \ln ((1 + \ 0.512\ ) / (1 - \ 0.512\ )) \\
&0.5 \ln ((\quad 1.512\ ) / (\quad 0.488\ )) \\
&0.5 \ln (\quad 3.09836\qquad\qquad\qquad\quad) \\
&0.5 \qquad\ 1.13087 \\
&\qquad\quad\ 0.56543
\end{aligned}
$$

The correlation (r=.9418) between FEFT total scores and scores on the 35 item Part A of the FEFT is expressed as Z as 1.75374.

The researcher using classical statistical procedures will presume that the standard error of Z=0.56543 and the standard error of Z=1.75374 are <u>exactly equal</u> (Glass & Hopkins, 1984, p. 306), i.e., .188982 (1/(n-3)**.5 = 1/28**.5 = .035714**.5). This presumption means that the researcher will assume that the 95% confidence intervals about both Zs are also exactly equal in their width, i.e.,

Z + and - 1.96 * .188982, or

Z + and - .3704047.

It seems illogical to make strong assumptions that standard errors are randomly and normally distributed, when one has data in hand that can be employed to <u>empirically</u> estimate standard error. "Bootstrap" methods (Efron, 1982, 1986) provide sophisticated estimates of the standard errors of results, informed by the data in hand rather than by paltry assumptions about the likely distribution of sample-estimates of parameters. Thus, Lunneborg (1987, p. 38, his emphasis) characterizes these as "<u>real</u>" estimates. And various types of confidence intervals can be constructed using these methods (Buckland, 1985; Efron, 1987; Efron & Tibshirani, 1986; Lunneborg, 1986).

Conceptually, "bootstrap" methods involve copying the data set over again and again many, many times into a large "mega" data set. Then dozens (or hundreds or thousands) of different samples are drawn from the "mega" file, and results are computed separately for each sample and then averaged. The method is powerful because the analysis considers so many configurations of subjects and informs the researcher regarding the extent to which results generalize across different configurations of subjects. Lunneborg (1987) has offered some excellent computer programs that automate this logic for univariate applications; Thompson (1988d) and Lambert, Wildt and Durand (1990) provide similar software for some multivariate applications. Borrello and Thompson (1989) and Scott, Thompson and Sexton (1989) illustrate applications of these methods.

Table 4 presents selected entries from bootstrap estimation of the correlation matrix presented in Table 1, based on 1,000 resamplings of the data from the 31 subjects. These analyses were conducted on a microcomputer using Lunneborg's (1987) program, CORBOOT.

INSERT TABLE 4 ABOUT HERE.

Lunneborg's (1987) program, BOOTLV, was run to derive various descriptive statistics for the results associated with the 1,000 resamplings. Some of these results are presented in Table 5. The tabled results suggest some important benefits of "bootstrap" estimation procedures. For example, the fact that the standard

deviation of the estimates over 1,000 resamplings (empirical estimates of standard error, of the r for the variable pair GEFT with FEFT Total (SD=.12698800) did not equal the SD for the variable pair FEFT Part A with FEFT Total (SD=.02987297) illustrates that it may not be tenable to assume that all standard errors for a fixed sample size are exactly equal. Furthermore, the fact that mean (e.g., .50920460) and median (e.g., .52025070) estimates of the r for the variable pair GEFT with FEFT Total were not exactly equal illustrates that the assumption that sampling error is normally distributed may also not always be exactly true.

INSERT TABLE 5 ABOUT HERE.

Lunneborg's (1987) program, BOOTCI, was run on the microcomputer to calculate 95% confidence intervals. Selected results for the example are presented in Table 6. The program will compute any width intervals the user desires, and also provides intervals constructed using several different logics.

INSERT TABLE 6 ABOUT HERE.

In the present example the conventional standard error for r expressed as Z was .188982 (1/(n-3)**.5 = 1/28**.5 = .035714**.5). The empirical estimates, i.e., the SD over 1,000 resamplings, was SD of r = .12698800; the Z form of this SE equals .12767. The smaller empirical estimate of the SE results in narrower 95% confidence intervals for the bootstrap procedure, as illustrated in Table 6.

To illustrate the invariance of bootstrap estimates (at least when 1,000 resamplings are conducted), a new set of 1,000 resamplings was isolated. Descriptive statistics for the analysis are presented in Table 7, and can be compared with the results reported in Table 5.

INSERT TABLE 7 ABOUT HERE.

Discussion

Three general comments can be made regarding the three effect size corrections reported here. First, the Wherry (1931) and the Herzberg (1969) corrections tend to be larger as either effects sizes or sample sizes become smaller, as illustrated by Thompson (1990). Thus, with a very large effect size approaching 1.0, or a large sample size, or both, it will matter less which, if any, statistical corrections the researcher applies in estimating effect sizes.

However, meta-analyses of substantive findings (effect sizes tend to be larger in reliability or criterion-related validity studies) suggest that effect sizes do not tend to get much larger than 25 to 33%. For example, Cohen's (1988) perusal of published research suggests that a correlation ratio of around 25% (r=.5) should be considered large in terms of typical findings across disciplines. The empirical meta-analytic work of Glass and others, which has yielded some additional ways of evaluating effect size, has also led to similar conclusions:

> In none of the dozen or so research literatures that

> we have integrated in the past five years have we ever encountered a cross-validated multiple correlation between study findings and study characteristics that was larger than approximately 0.60. That is, I haven't seen a body of literature in which we can account for much more than a third of the variability in the results of studies, [which is distinct from talking about results for only one smaller group of subjects]. (Glass, 1979, p. 13)

Second, as suggested by the illustration, the Herzberg (1969) correction tends to be more conservative than the Wherry (1931) correction. However, this only makes sense. Since the Herzberg correction is used to estimate effect sizes that may be isolated in a new sample, the estimate in effect must correct both for sampling error influences for the data in hand and for the sampling error that will recur in new samples. The Wherry correction only presumes one set of sampling error influences, i.e., sampling error for the data in hand.

Third, as illustrated by the case study, statistical estimates of standard errors for effect sizes and empirically derived estimates may differ, and the confidence intervals derived using the two approaches may also differ. It seems illogical to be prepared to accept the sample-based estimates, such as an r or a squared r, and to then to estimate the standard error of the sample-grounded estimates based on strong assumptions rather than the data in hand. Such illogic may have been necessary in the era

preceding both the elaboration of "bootstrap" logic (Efron, 1979, 1982, 1986, 1987; Efron & Tibshirani, 1986; Lunneborg, 1985, 1987, in press; Lunneborg & Tousignant, 1985) and the widespread availability of computer programs that readily implement the necessary calculations (e.g., Lunneborg, 1987; Thompson, 1988d). But a new day upon us.

Put differently, assumptions that sampling error is normally distributed might be tenable if researchers lived in a world in which they routinely drew true probability samples from defined populations, and if in this world all randomly selected subjects then participated in studies. But most of us inhabit a world in which samples of convenience may be necessary (though thoughtful usually compare the characteristics of the sample of convenience with those of the population), and a world in which some subjects do not agree to participate and still more withdraw subsequent to agreeing to participate.

Bootstrap logics are appealing, because they focus on the sina qua non of science, i.e., replication. As Thompson (1989b, p. 4) notes, "significance, importance, and replicability are all important issues in research. [But] Too many researchers attend only to issues of significance in their research. And in some respects, statistical significance may be the least important element of this research triumvirate."

References

Borg, W.R., & Gall, M.D. (1989). Educational research: An introduction (5th ed.). New York: Longman.

Borrello, G., & Thompson, B. (1989). A replication "bootstrap" analysis of the structure underlying perceptions of stereotypic love. Journal of General Psychology, 116, 317-327.

Buckland, S.T. (1985). Calculation of Monte Carlo confidence intervals. Applied Statistics, 34, 296-301.

Carter, D.S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

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

Chow, S.L. (1988). Significance test or effect size? Psychological Bulletin, 103(1), 105-110.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society (supplement), 4, 102-118.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14, 205-216.

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

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

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in

statistics. Scientific American, 248(5), 116-130.

Eason, S.H., & Daniel, L.G. (1989, January). Trends and methodological practices in several cohorts of dissertations. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 306 299)

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

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society of Industrial and Applied Mathematics, CBMS-NSF Monographs, 38.

Efron, B. (1986, April). On bootstrap inference. Paper presented at the annual meeting of the American Educational Research Association Meeting, San Francisco.

Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82, 171-185.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 1, 54-75.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver and Boyd.

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.

Glass, G.V, & Hopkins, K.D. (1984). Statistical methods in

education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

Guilford, J.P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.

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

Herzberg, P.A. (1969). The parameters of cross validation. Psychometrika, Monograph supplement, No. 16.

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

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

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

Kulik, J.A., & Kulik, C.L. (in press). Meta-analysis: Historical origins and contemporary practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2). Greenwich, CT: JAI Press.

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

Lambert, Z.V., Wildt, A.R., & Durand, R.M. (1990). Assessing sampling variation relative to number-of-factors criteria.

Educational and Psychological Measurement, 50, 33-48.

Lunneborg, C.E. (1985). Estimating the correlation coefficient: The bootstrap approach. Psychological Bulletin, 98, 209-215.

Lunneborg, C.E. (1986). Confidence intervals for a quantile contrast: Application of the bootstrap. Journal of Applied Psychology, 71, 451-456.

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences (Vol. 1). Seattle: University of Washington.

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

Lunneborg, C.E., & Tousignant, J.P. (1985). Efron's bootstrap with an application to the repeated measures design. Multivariate Behavioral Research, 20, 161-178.

Maxwell, S.E., Camp, C.J., & Arvey, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

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.

Melancon, J.G., & Thompson, B. (1987, November). Measurement characteristics of a test of field-independence: Literature review and development of the Finding Embedded Figures Test. Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile. (ERIC Document Reproduction Service No. ED 292 823)

Melancon, J. G., & Thompson, B. (1989). Measurement characteristics of the Finding Embedded Figures Test. Psychology in the Schools, 26, 69-78.

Melancon, J., & Thompson, B. (1990a, January). Latent trait calibrations for the Finding Embedded Figures Test: A study with middle school students. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. ED 314 498

Melancon, J., & Thompson, B. (1990b). Maximizing test reliability by stepwise variable deletion: A case study with the Finding Embedded Figures Test. Perceptual and Motor Skills, 70, 99-110.

Melancon, J., & Thompson, B. (1990c, January). Measurement characteristics of the Finding Embedded Figures Test in "speed" versus "power" administrations. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 314 495)

Olkin, I., & Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 29, 201-211.

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 51, 268-283.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills: Sage.

Rosnow, R.L., & Rosenthal, R. (1988). Focused tests of significance

and effect size estimation in counseling psychology. Journal of Counseling Psychology, 35, 203-208.

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

Scott, R.L., Thompson, B., & Sexton, D. (1989). Structure of a short form of the Questionnaire on Resources and Stress: A bootstrap factor analysis. Educational and Psychological Measurement, 49, 409-419.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Erlbaum.

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

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

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

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

Thompson, B. (1987b, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association,

Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

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

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

Thompson, B. (1988c). A note on statistical significance testing. Measurement and Evaluation in Counseling and Development, 20(4), 146-148.

Thompson, B. (1988d). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. Educational and Psychological Measurement, 48, 681-686.

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

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

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and

Development, 22, 2-6.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50, 15-31.

Thompson, B., & Melancon, J.G. (1990). Measurement characteristics of the Finding Embedded Figures Test: A comparison across three samples and two response formats. Educational and Psychological Measurement, 50, 333-342.

Wherry, R.J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-451.

Table 1
Bivariate Correlation Matrix
(n=31)

	FEFTA	FEFTB	FEFTTOT
FEFT Form A (y=35)			
FEFT Form B (y=35)	.7617		
FEFT Total (y=70)	.9418	.9352	
GEFT Total (y=18)	.4239	.5402	.5120

Table 2
Effect Size Estimates Invoking the Wherry Correction
(n=31)

Variables Being Correlated		r from Table 1	2 r	Adjusted r square
FEFT B	FEFT A	0.7617	0.580186	0.565710
FEFTTOT	FEFT A	0.9418	0.886987	0.883090
FEFTTOT	FEFT B	0.9352	0.874599	0.870274
GEFT	FEFT A	0.4239	0.179691	0.151404
GEFT	FEFT B	0.5402	0.291816	0.267395
GEFT	FEFTTOT	0.5120	0.262144	0.236700

Table 3
Effect Size Estimates Invoking the Herzberg Correction
(n=31)

Variables Being Correlated		r from Table 1	2 r	Adjusted r square
FEFT B	FEFT A	0.7617	0.580186	0.535690
FEFTTOT	FEFT A	0.9418	0.886987	0.875008
FEFTTOT	FEFT B	0.9352	0.874599	0.861307
GEFT	FEFT A	0.4239	0.179691	0.092746
GEFT	FEFT B	0.5402	0.291816	0.216755
GEFT	FEFTTOT	0.5120	0.262144	0.183938

Table 4
Calculated Correlation Coefficients for the Sample of 31 Subjects and Seven of 1,000 Random Resamplings of the 31 Subjects

Sample	Estimates of the Six Unique r's			
0	.42388050	.54021650	.51195450	.76173090
	.94184750	.93515300		
1	.44996280	.38785020	.45017010	.73693000
	.93473370	.92903820		
2	.40545540	.57462420	.54441740	.64641050
	.89779330	.91637680		
3	.26309840	.60893980	.48483890	.56424870
	.89433320	.87400320		
4	.69394230	.79731420	.77426050	.85819620
	.96291700	.96486490		
5	.50877910	.65551280	.61955510	.75999140
	.93972070	.93641970		
. . . .				
999	.57282050	.68778800	.66222770	.80693140
	.95198860	.94900570		
1000	.34975410	.40709220	.42246560	.59790740
	.89981100	.88771040		

Note. Sample 0 involves the results calculated for the 31 subjects. The correlation coefficients are presented in the order: (a) GEFT x FEFT Part A; (b) GEFT x FEFT Part B; (c) GEFT x FEFT total; (d) FEFT Part A x FEFT Part B; (e) FEFT total x FEFT Part A; and (f) FEFT total x FEFT Part B.

Table 5
Bootstrap Results Across 1,000 Resamplings of 31 Subjects in Random Configurations

Statistic	GEFT x FEFT Total	FEFT A x FEFT Total
r for 31 Subjects	.51195	.94185
Mean of 1,000 Samples	.50920460	.93533360
SD (akin to SE)	.12698800	.02987297
Median of 1,000 Samples	.52025070	.94288290
Lowest Estimate	-.006179	.7603
Largest Estimate	.8108	.9907
Range	.8170	.2305

Table 6
Conventional and Four Bootstrap Confidence Intervals
for the GEFT x FEFT Total Variable Pair

Conventional Estimates	
r	0.51195
r as *Z*	0.56536
Statistical SE	0.188982
95% CI for *Z*	0.194954 to 0.935765
Bootstrap Estimates	
r	0.51195
r as *Z*	0.56536
Empirical SE	0.12699
	Symmetric (Normal Theory)
95% CI for *r*	0.26263 to 0.76128
as *Z*	0.26893 to 0.99925
	Percentile Method
95% CI for *r*	0.23514 to 0.72782
as *Z*	0.23962 to 0.92407
	Bias Corrected Percentile
95% CI for *r*	0.21984 to 0.72028
as *Z*	0.22348 to 0.90822
	Minimum Width
95% CI for *r*	0.26930 to 0.75235
as *Z*	0.27610 to 0.97834

Table 7
Bootstrap Results A New 1,000 Resamplings
of 31 Subjects in Random Configurations

Statistic	GEFT x FEFT Total	FEFT A x FEFT Total
r for 31 Subjects	.51195	.94185
Mean of 1,000 Samples	.51206590	.93687050
SD (akin to SE)	.12820730	.02870789
Median of 1,000 Samples	.52079060	.94280670
Lowest Estimate	.06310	.8200
Largest Estimate	.8790	.9907
Range	.8159	.1706

APPENDIX A:
Item Analysis for Heuristic FEFT Data
(n=31)

Form A Item	p	SD	Corr. r with 39 FEFT A	Corr. r with 34 FEFT A	Corr. r with 17 GEFT
A 1	0.94	0.25	-0.01	0.05	0.49
A 2	0.84	0.37	0.33	0.38	0.28
A 4	0.81	0.40	-0.01	-0.08	-0.09
A 5	0.52	0.51	0.45	0.47	0.01
A 9	0.55	0.51	0.43	0.38	0.43
A11	0.81	0.40	0.25	0.23	0.17
A13	0.84	0.37	0.27	0.19	0.24
A17	0.77	0.43	0.27	0.29	0.10
A18	1.00	0.00	--	--	--
A23	0.94	0.25	0.35	0.38	0.31
A25	0.81	0.40	0.28	0.27	0.22
A26	0.68	0.48	0.33	0.31	0.09
A27	0.94	0.25	0.18	0.24	0.15
A28	0.90	0.30	0.05	0.04	-0.21
A30	0.81	0.40	0.30	0.27	0.12
A31	0.84	0.37	0.50	0.49	0.26
A32	0.87	0.34	0.40	0.35	0.09
A33	0.87	0.34	0.40	0.37	0.01
A34	0.84	0.37	0.66	0.61	0.03
A35	0.87	0.34	0.42	0.35	0.03
A36	0.65	0.49	0.42	--	0.15
A37	0.71	0.46	0.09	--	-0.18
A38	0.32	0.48	0.27	--	-0.08
A39	0.74	0.44	0.43	--	0.22
A40	0.52	0.51	0.32	--	0.29
A 3 L01	0.97	0.18	0.03	0.03	-0.02
A 6 L02	0.84	0.37	0.25	0.26	0.45
A 7 L03	0.35	0.49	0.07	0.06	0.33
A 8 L04	0.71	0.46	0.64	0.62	0.26
A10 L05	0.52	0.51	0.29	0.28	-0.04
A12 L06	0.61	0.50	0.27	0.30	0.18
A14 L07	0.90	0.30	0.21	0.25	0.24
A15 L08	0.48	0.51	0.75	0.76	0.45
A16 L09	0.87	0.34	0.19	0.12	0.09
A19 L10	0.87	0.34	0.24	0.28	0.01
A20 L11	0.77	0.43	0.27	0.27	0.18
A21 L12	0.94	0.25	0.16	0.16	-0.11
A22 L13	0.97	0.18	0.45	0.50	0.23
A24 L14	0.87	0.34	0.26	0.30	0.05
A29 L15	0.77	0.43	0.55	0.57	0.30
Non-Linking					
Mean	0.774	0.379	0.306	0.294	0.131
SD	0.157	0.109	0.157	0.160	0.172

Linking					
Mean	0.763	0.374	0.307	0.319	0.174
SD	0.183	0.108	0.196	0.201	0.169
All Form A					
Mean	0.770	0.377	0.299	0.305	0.144
SD	0.167	0.109	0.178	0.180	0.171

Form B Item	p	SD	Corr. r with 39 FEFT B	Corr. r with 34 FEFT B	Corr. r with 17 GEFT
B 2	0.42	0.50	0.40	0.38	0.38
B 3	0.26	0.44	0.21	0.21	0.22
B 4	0.55	0.51	0.15	0.18	0.04
B 7	0.81	0.40	0.30	0.31	-0.09
B 8	0.81	0.40	0.23	0.20	0.12
B10	0.74	0.44	0.33	0.34	0.15
B12	0.77	0.43	0.43	0.46	0.21
B13	0.23	0.43	0.38	0.31	0.43
B16	0.74	0.44	0.32	0.36	0.19
B19	1.00	0.00	--	--	--
B20	0.94	0.25	0.24	0.31	0.23
B23	0.87	0.34	0.20	0.23	0.16
B24	0.81	0.40	0.39	0.43	0.18
B25	0.81	0.40	0.44	0.43	0.31
B27	1.00	0.00	--	--	--
B30	0.55	0.51	0.35	0.42	0.16
B31	0.94	0.25	0.43	0.46	0.15
B33	0.90	0.30	0.26	0.21	0.13
B34	1.00	0.00	--	--	--
B35	0.81	0.40	0.34	0.33	0.22
B36	0.19	0.40	0.21	--	0.25
B37	0.16	0.37	0.11	--	0.23
B38	0.42	0.50	0.15	--	0.34
B39	0.42	0.50	0.28	--	0.35
B40	0.52	0.51	0.12	--	0.25
B 1 L01	0.97	0.18	0.50	0.48	0.23
B 5 L02	0.81	0.40	0.40	0.39	0.18
B 6 L03	0.39	0.50	-0.02	-0.02	0.12
B 9 L04	0.77	0.43	0.24	0.26	0.20
B11 L05	0.68	0.48	0.29	0.30	0.23
B14 L06	0.74	0.44	0.24	0.22	0.29
B15 L07	0.90	0.30	0.30	0.26	0.24
B17 L08	0.48	0.51	0.23	0.22	0.26
B18 L09	0.90	0.30	0.28	0.33	0.35
B21 L10	0.87	0.34	0.48	0.50	0.30
B22 L11	0.81	0.40	0.30	0.22	0.10
B26 L12	0.94	0.25	0.13	0.14	-0.06
B28 L13	0.94	0.25	0.48	0.46	0.20
B29 L14	0.84	0.37	-0.01	-0.04	0.26
B32 L15	0.84	0.37	0.51	0.52	0.35

Non-Linking					
Mean	0.666	0.365	0.318	0.328	0.210
SD	0.266	0.153	0.086	0.092	0.112
Linking					
Mean	0.791	0.368	0.291	0.283	0.218
SD	0.160	0.094	0.163	0.167	0.101
All Form B					
Mean	0.713	0.366	0.266	0.306	0.197
SD	0.240	0.134	0.146	0.134	0.118

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 29, 1991