DOCUMENT RESUME

ED 344 895 TM 018 174

AUTHOR Thompson, Bruce
TITLE **Exploring the Replicability of a Study's Results:** Bootstrap Statistics for the Multivariate Case.
PUB DATE 8 Apr 92
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Correlation; Discriminant Analysis; *Multivariate Analysis; Regression (Statistics); *Research Methodology; Sampling; *Statistical Significance
IDENTIFIERS *Bootstrap Methods; Empirical Research; Linear Models; *Research Replication; T Test

ABSTRACT

Conventional statistical significance tests do not inform the researcher regarding the likelihood that results will replicate. One strategy for evaluating result replication is to use a "bootstrap" resampling of a study's data so that the stability of results across numerous configurations of the subjects can be explored. This paper illustrates the use of the bootstrap in a canonical correlation analysis. Canonical correlation analysis is the most general case of classical general linear model analyses, subsuming other univariate and multivariate parametric method (e.g., t-tests, analysis of variance, analysis of covariance, regression, multivariate analysis of variance, and discriminant analysis) as special cases. A sample of 50 out of 301 subjects from a study by K. J. Holzinger and F. Swineford (1939) is used. Since bootstrap analyses capitalize during resampling on the commonalities inherent in a given sample, they yield somewhat inflated evaluations of replicability. However, inflated empirical evaluations of replicability are often superior to a mere presumption of replicability. Ten tables and one figure present details of the analysis. A 63-item list of references and an appendix listing the 50 analysis cases are included. (Author/SLD)

canstrap.wpl 4/8/92

Exploring the Replicability of a Study's Results: Bootstrap Statistics for the Multivariate Case

Bruce Thompson

Texas A&M University 77843-4225

and

Baylor College of Medicine

Paper presented at the annual meeting (session #28.50) of the American Educational Research Association, San Francisco, CA, April 22, 1992.

ABSTRACT

Conventional statistical significance tests do not inform the researcher regarding the likelihood that results will replicate. One strategy for evaluating result replicability is to employ a "bootstrap" resampling of a study's data so that the stability of results across numerous configurations of the subjects can be explored. The present paper illustrates the use of the bootstrap in a canonical correlation analysis. Canonical correlation analysis is the most general case of classical general linear model analyses, subsuming other univariate and multivariate parametric methods (e.g., _t_-tests, ANOVA, ANCOVA, _r_, regression, MANOVA, and discriminant analysis) as special cases.

The use of statistical significance testing as part of the interpretation of empirical research results has historically generated considerable debate (Carver, 1978; Huberty, 1987; Morrison & Henkel, 1970; Thompson, 1989a, 1989c, 1989e). A series of articles on the limits of statistical significance testing has even appeared on a seemingly periodic basis in recent editions of the American Psychologist (Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989). Thompson (1992c) points out several of the many possible objections to overreliance on conventional statistical significance testing. Two of these objections are most noteworthy.

1. Statistical Significance Testing can be Tautological

Even some widely respected authors of prominent methodology textbooks at times take internally inconsistent positions with respect to the role that conventional statistical significance testing should play in analysis (see book reviews by Thompson, 1987a, 1988d). And some dissertation authors may be disproportionately susceptible to excessive awe for significance tests (LaGaccia, 1991; Thompson, 1988b). But researchers who have had the experience of working with large samples (cf. Kaiser, 1976) soon realize that virtually all null hypotheses will be rejected at some sample size, 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." Many researchers possess this insight, but somehow *do not integrate this knowledge into their paradigms* for conceptualizing or conducting research. Thus, the insight too rarely impacts actual practice.

Although statistical significance is a function of at least seven interrelated features of a study (Schneider & Darcy, 1984), sample size is a basic influence on significance. To some extent significance tests evaluate the size of the researcher's sample--most researchers already know prior to conducting significance tests whether the sample in hand is large or small, so these outcomes do not always yield understanding that would be lost absent a significance test. As Thompson (1992b, p. 436) notes:

> 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. This tautology has created considerable damage as regards the cumulation of knowledge...

2. <u>Sole Reliance on Statistical Significance Testing Creates Inescapable Dilemmas for Researchers</u>

Researchers who place an inordinate emphasis on statistical significance tests also often confront an inescapable dilemma, though most researchers do not recognize (or prefer to ignore) this dilemma. All statistical significance tests invoke certain

assumptions. For example, ANOVA requires pooling the variances of the dependent variable across the cells of the design during the calculation of the mean square used in the denominator of the fixed-effects F-test (Haase & Thompson, 1992). This pooling is legitimate if and only if the variances of the dependent variable scores in all the cells are essentially equal. This is the well known "homogeneity (i.e., equality) of variance" assumption.

Similarly, as Thompson (1992a) notes, ANCOVA is a three-stage analysis in which (a) regression weights for the covariate are derived *completely ignoring* group or cell membership of the subjects, (b) predicted dependent variable scores ($\hat{Y}$) are computed using the weights, and are then subtracted from the actual dependent variable scores (Y) of the subjects to yield an "e" score ("e_i" = $Y_i - \hat{Y}_i$) for each ith subject, and then (c) an ANOVA is conducted using the "e" scores as the dependent variable in place of the Y scores. As Loftin and Madison (1991) explain in some detail, this process is legitimate if and only if the regression equations for predicting Y with the covariate(s) are essentially the same, i.e., the "homogeneity of regression" assumption is met. Because a single regression equation, a single equation that is calculated completely ignoring group membership, is employed to statistically adjust the Y scores, this single equation can only reasonably be used if the equations for the different groups or cells are reasonably comparable, otherwise use of a "pooled" regression equation would be inappropriate.

Many researchers use statistical significance testing to

evaluate *both* their preliminary methodological assumption hypotheses (e.g., the ANOVA homogeneity of variance assumption, the ANCOVA homogeneity of regression assumption) and their substantive hypotheses (e.g., the mean dependent variable score of the treatment group equals that of the control group). These researchers hope to <u>not</u> reject the null hypotheses involving methodological assumptions (e.g., they want the dependent variable variances in the cells to all be equal), while they typically hope to reject their substantive hypotheses. But as Thompson (1991c, p. 504) notes, this creates a dilemma, since

> the same large sample size that yields power against Type II error in testing the substantive hypotheses of interest in ANCOVA [or ANOVA or the <u>t</u>-test] is also going to tend to yield statistically significant effects for the preliminary homogeneity of regression [or of variance] test.

Some researchers attempt to escape this dilemma by presuming that their methods are robust to the violation of their assumptions. This does not generally appear to be the case with respect to ANCOVA (Keppel & Zedeck, 1989). And the longstanding view that ANOVA was robust to the violation of the homogeneity of variance assumption has recently been called into some question, thanks to more sophisticated Monte Carlo studies conducted with more complicated designs, and with more simulation samples (e.g., Rogan & Keselman, 1977; Tomarkin & Serlin, 1986; Wilcox, Charlin & Thompson, 1986).

Alternatives

Over the years various alternatives that might serve as substitutes for or augmentations of statistical significance testing have been proposed. For example, Serlin and Lapsley (1985) advocated placing an emphasis on confidence intervals, Bayesian approaches have been encouraged by some (e.g., Good, 1981), and somewhat less serious proposals have been presented by still others (Salzman, 1989).

But some strategies emphasize interpretation based on the estimated likelihood that results will replicate. This emphasis is compatible with the basic purpose of science: isolating conclusions that replicate under stated conditions. Notwithstanding some misconceptions to the contrary, conventional statistical significance tests do not evaluate the probability that results will generalize (Carver, 1978).

A particularly powerful strategy for evaluating result replicability invokes the *bootstrap* methods developed by Efron and his colleagues (cf. Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1990). Conceptually, these methods involve copying the data set over again and again many many times into an infinitely large "mega" data set. Then 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 (including configurations in which a subject may be represented several times or not at all) and informs

the researcher regarding the extent to which results generalize across different types of subjects. Lunneborg (1987) has offered some excellent computer programs that automate this logic for univariate applications; Thompson (1988) provides similar software for multivariate applications. Recently, user-friendly PC bootstrap software has become available from publishers around the world, e.g., the menu-driven program, BOJA, distributed by iecProGAMMA, P.O. Box 841, 9700 AV Groningen, The Netherlands.[1]

Bootstrap versus Classical Statistical Significance Tests

All statistical tests invoke four estimates. The first is a *single* statistic estimating a single population parameter calculated from the sample data in hand. The remaining three estimates are calculated <u>not</u> from the data in hand, but rather from entirely different data (i.e., the sampling distribution of the estimated parameter) conceptually involving *multiple* repeated samplings of the parameter estimate from a population. These four estimates are: (a) the single parameter estimate (e.g., $\bar{X}$, r) derived from a sample believed to be representative of a population; (b) the second moment about the mean of multiple estimates of the parameter of interest (i.e., the standard deviation (SD) of the repeatedly sampled estimates--the standard error (SE_E) of the estimated statistic); (c) the third moment about the mean of multiple estimates of the parameter (i.e., the coefficient of $skewness_E$); and (d) the fourth moment about the mean of multiple estimates of the parameter (i.e., the coefficient of $kurtosis_E$).

Many researchers recognize the use of the first two statistics in their analyses. Thus, researchers using LISREL and EQS analyses routinely pay more attention to parameter estimates that are greater than the individual standard errors of given estimates. As Kerlinger (1986, chapter 12) explains in some detail, test statistics also invoke the ratio of a parameter estimate to the SE_E. For example, researchers often use a t-test to evaluate the null hypothesis that a mean equals zero. For a sample of size n, the SD of infinitely many samples of size n from a population in which the mean is zero (i.e., $SE_{\bar{X}}$) would be approximately $SD_X/(n^{**}.5)$. The test statistic, t, for this research situation is calculated as the ratio, $\bar{X} / (SD_X/(n^{**}.5))$.

The use of the third and fourth statistics is not so explicit. But when we evaluate the probability of our sample result, $p_{CALCULATED}$ given an assumption that the null is true, we usually compare our result against the α (or the $\alpha/2$) percentile of the test statistic, and the skewness and the kurtosis of this sampling distribution are part of what dictates what will be the value the α%ile of the test distribution. Of course, conventional confidence intervals employ exactly the same elements as statistical significance testing, and do make the use of all four estimates explicitly obvious (Glass & Hopkins, 1984, section 11.7).

However, *it is contradictory to be willing to use the sample to derive our (a) parameter estimate, and to be unwilling to let the sample offer similar insight regarding the (b) SE of our estimate, and regarding the (c) skewness and (d) kurtosis of*

sampled estimates. One way to let our data speak regarding the latter three estimates is to conduct a bootstrap analysis, i.e., we momentarily treat our sample data as if it constituted the population and we draw numerous (usually at least a thousand) random samples from the sample to infer what the sampling distribution looks like. To mimic randomly sampling our data with n subjects from the population, we do all our "resampling" from our mock population by drawing random samples with replacement from our data in hand, and to honor our research situation each resample is drawn to also have exactly size n.

The bootstrap approach can also be employed to yield a variety of confidence intervals, which vary as a function of the assumptions they make about the sampling distribution. Of course, bootstrap and other methods that focus on the invariance or the generalizability of results are no more magical than is classical statistical significance testing itself. No analytic methods can take us beyond the limits of our data. We use methods to let data speak in various ways, not to make data more than they can be.

A Bootstrap Example for the Univariate Case

The Table 1 data can be used to illustrate a bootstrap application and its potential benefits. These estimates were developed using the software available from Lunneborg (1987), and were based on 1,000 samples with replacement. As reported in Table 2, the standard deviation of the 1,000 estimates of r was .173--this is the *empirical* estimate of SE_r, and is considerably smaller than the estimate of the SE (SE_{Z_r} = .354, SE_r = .339) derived based

on assumptions.[2] Figure 1 graphically presents the bootstrap results. The bootstrap results were also useful in alerting the researcher to the fact that the sampling distribution may not be normal, e.g., the distribution may be negatively skewed.

INSERT TABLES 1 AND 2 AND FIGURE 1 ABOUT HERE.

The bootstrap approach can be employed to yield a variety of confidence intervals, which vary as a function of the assumptions they make about the sampling distribution. The three estimates calculated by the Lunneborg (1987) program for the Table 1 data are reported in Table 2. The "bias corrected" estimate makes the fewest assumptions regarding the sampling distribution (Lunneborg, 1987, p. 54), that is, relies most upon the empirical findings from resampling. Since none of the confidence intervals subsume zero, the bootstrap results employing an empirically estimated sampling distribution, unlike the conventional approach, yields a statistically significant result.

Bootstrap Multivariate Methods

Most of the previous bootstrap software applications have been implemented in univariate statistical applications. However, it might be argued that such methods would be even more useful in the multivariate case, since in theory multivariate methods offer even more opportunities to capitalize on sampling error (e.g., Gorsuch, 1983, p. 330).

The major barrier to conducting a multivariate bootstrap involves the multidimensional character of the "space" in which the

analysis is conducted. The bootstrap must be applied such that each of the hundreds or thousands of resampling results are all located in a common factor space before the mean, SD, skewness and kurtosis are computed.

For example, in a factor analysis of population data, the first two principal components of IQ data might be "Verbal" and "Performance", and the eigenvalues of the two factors prior to rotation (Thompson, 1989d) might be 5.5 and 5.4, respectively. In various samples from this population the two components might emerge very much as the same constructs, but sampling error might introduce small variations in the ordering of the two factors within the analysis, with "Verbal" being the first factor in some solutions but the second factor in other samples.

If the analyst computed mean structure (or pattern) coefficients for the first variable on the first component across all the repeated samplings, the mean would be a nonsensical mess representing an average of some apples, some oranges, and perhaps some kiwi. The sampled solutions must be rotated to best fit positions with a common target solution, prior to computing means and other statistics across the samples, so that the results are reasonable.

The same considerations apply when one is considering resampling from sample data in a bootstrap analysis, as against a meta-analysis of independent samples from a population (e.g., Thompson, 1989b). Several viable candidates for the target used to define to common factor space that links results across resamplings

can be identified. These include:

(a) a matrix of zeroes, ones, and negative ones, defining a simple structure, delineated based on theory;

(b) a structure or a weight matrix isolated in previous research; or

(c) a structure or a weight matrix for the sample data in hand.

Bootstrap in the Canonical Case

The theoretical and the programming difficulties inherent in conducting bootstrap analyses with multivariate procedures have been overcome as regards factor analysis (Daniel, 1992; Lambert, Wildt & Durand, 1990; Thompson, 1988c) and discriminant analysis/one-way MANOVA (Lawson & Snyder, 1992). This work is noteworthy, since multivariate methods are often vitally important in social science research (Fish, 1988).

Thompson (1986, p. 9) notes that the reality about which most researchers wish to generalize is usually one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects." Tatsuoka's (1973, p. 273) 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.

Fish (1988) and Maxwell (in press) both present data illustrating how univariate and multivariate analysis of the same data can lead to radically different conclusions.

Although the availability of bootstrap software for factor analysis and for discriminant analysis/one-way MANOVA is helpful, it would also be useful to be able to bootstrap a canonical correlation analysis. Canonical correlation analysis is the most general case of classical general linear model analyses, subsuming other univariate and multivariate parametric methods (e.g., _t_-tests, ANOVA, ANCOVA, _r_, regression, MANOVA, and discriminant analysis) as special cases (Knapp, 1978; Xitao, 1992). Thompson (1988a, 1991b) illustrates these connections using small heuristic data sets to make the discussion concrete and accessible.

The present paper uses data from Holzinger and Swineford (1939, pp. 81-91) for heuristic purposes to illustrate a bootstrap of a canonical analysis. These cognitive ability data are widely available, and have been employed by many authors for similar illustrative purposes (e.g., Gorsuch, 1983, passim; Jöreskog & Sörbom, 1989, passim).

The heuristic example assumes two criterion variables, General Verbal Ability and Paragraph Comprehension scores, and four predictor variables: Speeded Dot Counting, Speeded Discrimination of Straight and Curved Capitals, Math Number Series, and Woody-

McCall Mixed Math Fundamentals scores. Table 3 presents the correlation matrix associated with the full data set (N=301).

INSERT TABLE 3 ABOUT HERE.

Canonical analysis partitions the correlation matrix into two "intradomain" and two "interdomain" quadrants. These four submatrices are then manipulated (see Thompson, 1984, pp. 11-16 for details) to yield a "quadruple-product matrix". The quadruple product matrix for these data is presented in Table 4.

INSERT TABLE 4 ABOUT HERE.

The quadruple-product matrix is then subjected to a principal components analysis. The eigenvalues of the quadruple-product are the squared canonical correlation coefficients (R_C^2). The number of squared canonical correlation coefficients always equals the number of variables in the smaller of the two variable sets, because that is the rank of the "quadruple-product" matrix.

Since conventional parametric methods are all correlational least squares analyses, all such analyses involve weights similar to the beta weights generated in regression. These weights 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 parametric methods, and to confuse graduate students.

All parametric methods also involve the creation of latent or synthetic variables analogous to the predicted dependent variable in regression ($\hat{Y}$). And all analyses can invoke the correlation coefficients between an observed and a latent variable (called a "structure correlation" or a "structure coefficient") an important aids to interpretation (Thompson & Borrello, 1985). Table 5 presents the canonical function, structure, correlation, and other coefficients associated with the canonical analysis of the Table 3 matrix.

INSERT TABLE 5 ABOUT HERE.

Table 6 presents the correlation coefficients for the same variables for a random sample (see Appendix A) of 50 of the 301 subjects in the population. Table 7 presents the canonical analysis of these sample data.

INSERT TABLES 6 AND 7 ABOUT HERE.

Program CANSTRAP (Thompson, in press) was then employed to resample 1,000 samples, each of size 50, from the Appendix A data. The resampling procedure in bootstrap typically invokes resamples of the same size as the sample itself, to mimic the influences on the actual sample size.

For this heuristic example the random resampling involved a mean use of the 50 subjects of 1,000 times each (SD=26.27). The smallest number of times a subject was drawn across 1,000 samples was 942. The most times a subject was drawn over 1,000 samples was

1,056.

Of course, since resampling is done with replacement, a given subject may be drawn more than once in a given resampling, or not at all. For example, in these analyses subject 8 from Appendix A was drawn twice in the first resampling, but subject 2 was not drawn at all in this resampling. However, in the second of the 1,000 resamplings, subject 8 was not drawn at all, but subject 9 was drawn three times.

Table 8 presents descriptive statistics for the squared canonical correlation coefficients for both functions I and II across the 1,000 resamplings. Table 9 presents descriptive statistics for the function and structure coefficients for the smaller variable set, computed only after first invoking a Procrustean rotation of each resampled function coefficient matrix to a best fit position with the Table 7 function coefficient matrix.[3] Table 10 presents the corresponding descriptive statistics for the function and structure coefficients for the larger variable set, across 1,000 resamplings.

INSERT TABLES 8, 9 AND 10 ABOUT HERE.

Discussion

With respect to bootstrap canonical effect sizes, the mean R_C^2 for function I across 1,000 resamplings was 23.703% (SD=.11744), as against the true population value of 29.982%, and the initial sample value of 35.574%. The standard deviation (.11744) about this mean estimate is an *empirical* estimate of the standard error

of the statistic. And the remaining moments about the mean advise the researcher that the resampled estimates are not normally distributed, as might be otherwise expected. Indeed, both sets of estimates are positively skewed, as reported in Table 8.

The finding that the canonical correlation coefficients are somewhat positively biased is fully expected, just as "shrinkage" dynamics are expected in regression effects. Indeed, it may be useful to invoke the same "shrinkage" corrections employed in regression (Fisk, 1991) with the resampled canonical estimates (Thompson, 1990).

In any case, the standard deviations of the resampled canonical correlations, akin to standard errors, should be carefully considered. For example, in the present study the mean R_c^2 for function I across 1,000 resamplings (43.958%) was within one SE (43.958% - 11.744% = 32.214%) of the actual sample result, i.e., 35.374%. And the resampled result (43.958%) was within two SEs of the actual population (29.982%) result in the example analog to a true population.

The standard errors for the function and structure coefficients, presented in Tables 9 and 10, indicate that both function and structure coefficients are highly susceptible to sampling error. Again, this result is consistent with previous Monte Carlo research (Thompson, 1991a). Such results alert the researcher to exercise considerable caution when interpreting canonical weights and structure coefficients.

In summary, the business of science is formulating

generalizable insight. No one study, taken singly, establishes the basis for such insight. As Neale and Liebert (1986, p. 290) observe:

> No one study, however shrewdly designed and carefully executed, can provide convincing support for a causal hypothesis or theoretical statement... Too many possible (if not plausible) confounds, limitations on generality, and alternative interpretations can be offered for any one observation. Moreover, each of the basic methods of research (experimental, correlational, and case study) and techniques of comparison (within- or between-subjects) has intrinsic limitations. How, then, does social science theory advance through research? The answer is, by collecting a diverse body of evidence about any major theoretical proposition.

Evaluating the generalizability of canonical results is a daunting task, but a task which the serious scholar can ill-afford to shirk. Such evaluations are important. As Nunnally (1978, p. 298) notes, "one tends to take advantage of chance in any situation [all parametric methods] where something is optimized from the data at hand", as in least squares methods, i.e., all conventional parametric methods.

Bootstrap analyses are one vehicle, but an important vehicle, for evaluating the replicability of results. The researcher may

vest more confidence in results that replicate over the numerous configurations of subjects created during a bootstrap analysis. Since such analyses capitalize during resampling on the commonalities inherent in a given sample in hand (e.g., measurement at a given point in time, perhaps measurement in a given geographic location), such analyses always yield somewhat inflated evaluations of replicability. But inflated empirical evaluations of replicability are often superior to a mere presumption of replicability, especially when the researcher can take this capitalization into account during interpretation.

References

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

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Daniel, L. (1992, April). Bootstrap methods in the principal components case. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED forthcoming)

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

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

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

Fisk, Y.H. (1991, April). Various approaches to effect size estimation. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED forthcoming)

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

Good, I.J. (1981). Some logic and history of hypothesis testing. In J. Pitt (Ed.,), Philosophy in economics (pp. 149-174). Dordrecht, Holland: Reidel.

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

Haase, T., & Thompson, B. (1992, January). The homogeneity of variance assumption in ANOVA: What it is and why it is required. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston.

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.

Holzinger, K.J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-polar solution (No. 48). Chicago, IL: University of Chicago. (data on pp. 81-91)

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

Jöreskog, K.G., & Sörbom, D. (1989). LISREL 7: A guide to the program and applications (2nd ed.). Chicago: SPSS.

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

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

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

Knapp, T. R. (1978). Canonical correlation analysis: A general

parametric significance testing system. Psychological Bulletin, 85, 410-416.

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

LaGaccia, S.S. (1991). Methodology choices in a cohort of education dissertations. In B. Thompson (Ed.) Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 149-158). Greenwich, CT: JAI Press.

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.

Lawson, S., & Snyder, P. (1992, April). Use of the bootstrap in discriminant function analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED forthcoming)

Loftin, L.B., & Madison, S.Q. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.

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

Lunneborg, C.E. (1990). [Review of Computer intensive methods for testing hypotheses]. Educational and Psychological Measurement, 50, 441-445.

Maxwell, S. (in press). Recent developments in MANOVA applications.

In B. Thompson (Ed.), Advances in social science methodology (Vol. 2). Greenwich, CT: JAI Press.

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.

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

Neale, J.M., & Liebert, R.M. (1986). Science and behavior: An introduction to methods of research (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J.C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Rogan, J.C., & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Association, 14, 493-498.

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

Salzman, K.L. (1989). A significantly significant approach to significant research findings: The Salzman All-Significant F test. In G.C. Ellenbogen (Ed.), The primal whimper (pp. 158-162). New York, NY: Guilford Press.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8, 573-582.

Serlin, R.C., & Lapsley, D. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40, 73-83.

Tatsuoka, M. M. (1973). 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. Menlo Park, CA: SAGE.

Thompson, B. (1986, November). Two reasons why multivariate methods are usually vital: An understandable reminder with concrete examples. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.

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

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

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

Thompson, B. (1989b). Meta-analysis of factor structure studies: A case study example with Bem's androgyny measure. Journal of Experimental Education, 57, 187-197.

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

Thompson, B. (1989d). Prerotation and postrotation eigenvalues shouldn't be confused: A reminder. Measurement and Evaluation in Counseling and Development, 22(3), 114-116.

Thompson, B. (1989e). 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. (1991a). Invariance of multivariate results. Journal of Experimental Education, 59, 367-382.

Thompson, B. (1991b). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1991c). [Review of Data analysis for research designs]. Educational and Psychological Measurement, 51, 500-510.

Thompson, B. (1992a). Misuse of ANCOVA and related "statistical control" procedures. Reading Psychology, 13, iii-xviii.

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

Thompson, B. (1992c, April). The use of statistical significance tests in research: Some criticisms and alternatives. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Thompson, B. (in press). CANSTRAP: A program that performs a bootstrap canonical correlation analysis. College Station, TX: Psychometrics Group.

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

Tomarkin, A.J., & Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 88, 90-99.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F' statistics. Communications and Statistics, 15, 933-943.

Xitao, Fan. (1992, April). Canonical correlation analysis as a general analytic model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED forthcoming)

Footnotes

[1]Examples of such software and the distributors of the software include: (a) "Resampling Stats", distributed by Resampling Stats, 612 N. Jackson, Arlington, VA 22201; (b) "Statistical Calculator", distributed by Erlbaum, 27 Palmeira Mansions, Church Road, Hove East Sussex BN3 2FA, United Kingdom; (c) SPIDA, distributed on behalf of its Australian author by SERC, 1107 NE 45th--Suite 520, Seattle, WA 98105; and (d) the menu-driven program, BOJA, distributed by iecProGAMMA, P.O. Box 841, 9700 AV Groningen, The Netherlands.

[2]Strictly speaking, the standard error (SE) of Z_r is only 1/((n-3)**.5) when the population $\underline{r}$ is zero. Thus, it is actually contradictory to calculate SE_{Z_r} based on an assumption that $\underline{r} = 0$, and to then use SE_{Z_r} to calculate confidence intervals for $\underline{r} \neq 0$, unless one only wishes to test H_0: $\underline{r} = 0$. In this case conceptually the CI is really being constructed around 0 (and not $\underline{r}$), and the test is whether the point estimate, $\underline{r}$, falls within the interval. However, in practice we usually consider this estimation procedure to be "close enough".

[3]Another viable candidate for the target matrix used to define a common factor space would be the eigenvector matrix of the quadruple product matrix.

Table 1
Hypothetical Data Used to Illustrate Bootstrap Evaluation of an Estimate of r

ID	Y	X
1	.18	.20
2	.54	1.88
3	-.49	-.76
4	.92	.42
5	.22	.32
6	.75	-.56
7	.66	1.55
8	-2.65	-1.21
9	-.51	-.66
10	.47	-.96
11	-.09	-.21
r_{YX}	.560	
Z_r	.632	

Note. $Z_r = 1.1513\ (\ln\ ((1 + |r|) / (1 - |r|)))$
$1.1513\ (\ln\ ((1 + .560) / (1 - .560)))$
$1.1513\ (\ln\ (1.560 / .440))$
$1.1513\ (\ln\ (3.541))$
$1.1513\ (.549) = .632$

Table 2
Conventional and Bootstrap Significance Tests
for $\underline{r}$=.560 for the Table 1 Data

Sampling Statistics/ Significance Tests	Classical Estimates Based on *Statistical Assumptions*	*Empirically* Based Bootstrap Estimates
Second Moment of the Sampling Distribution		
SE_{Z_r}	.354[a]	
SE_r	.339[b]	.173
Third Moment of the Sampling Distribution		
Coefficient of Skewness of $\underline{r}$	.000 (*assumed*)	-.780
Third Moment of the Sampling Distribution		
Coefficient of Kurtosis of $\underline{r}$	.000 (*assumed*)	1.895
Density of the Sampling Distribution		
90.0%ile of Z_r	1.282 (*assumed*)	1.037
95.0%ile of Z_r	1.645 (*assumed*)	1.164
97.5%ile of Z_r	1.960 (*assumed*)	1.324
95% Confidence Intervals		
About Z_r	-.061 to 1.325[c]	
About $\underline{r}$	-.060 to 0.868[d]	+.220 to +.899[e] +.188 to +.868[f] +.082 to +.822[g]

[a]*Calculated* as SE_{Z_r} = 1 / ((n - 3) ** .5) = 1 / ((11 - 3) ** .5) =
1 / (8 ** .5) = 1 / 2.828 = .354.
[b]*Calculated* as SE_{Z_r} = .354 converted back into SE_r = .339.
[c]*Calculated as* $CI_{95\%}$ about $Z_r = Z_r - (1.960 * SE_{Z_r})$ to $Z_r + (1.960 * SE_{Z_r})$
= .632 - (1.960 * .354) to .632 + (1.960 * .354)
= .632 - .693 to .632 + .693
[d]The conversion of $\underline{r}$ expressed as Fisher's Z transform back into $\underline{r}$.
[e]$CI_{95\%}$ calculated using symmetric or normal theory approach.
[f]$CI_{95\%}$ calculated using percentile method.
[g]$CI_{95\%}$ calculated using bias corrected method.

Table 3
Correlation Coefficients for "Population" of N=301
Subjects from the Holzinger and Swineford (1939) Study

		T5	T6	T12	T13	T23	T24
General Verbal	T5	1.0000	.6572	.1649	.2052	.3950	.3933
Paragraph Comprehension	T6	.6572	1.0000	.1069	.2078	.4516	.4353
Speeded Dot Counting	T12	.1649	.1069	1.0000	.4490	.2615	.3111
Speeded Discrimination Caps	T13	.2052	.2078	.4490	1.0000	.3322	.2824
Math Number Series	T23	.3950	.4516	.2615	.3322	1.0000	.4600
Woody-McCall Math	T24	.3933	.4353	.3111	.2824	.4600	1.0000

Table 4
"Quadruple-Product" Matrix Analyzed in Canonical Analysis
(N=301)

	1	2
1	.170	.093
2	.160	.273

Table 5
Canonical Function and Structure Coefficients for Population Data
(N=301)

	Function I			Function II			
Variable/ Coef.	Function	Structure	Squared Structure	Function	Structure	Squared Structure	h^2
T5	0.36902	0.84089	70.710%	-1.27437	-0.54120	29.290%	100.00%
T6	0.71803	0.96054	92.264%	1.11563	0.27814	7.736%	100.00%
Adequacy			81.487%			18.513%	
Redundancy			24.431%			0.165%	
Rc2			29.982%			0.893%	
Redundancy			12.498%			0.234%	
Adequacy			41.686%			26.252%	
T12	-0.13193	0.25128	6.314%	-1.06970	-0.96136	92.421%	98.74%
T13	0.11157	0.41088	16.882%	0.07567	-0.31409	9.865%	26.75%
T23	0.59236	0.85842	73.688%	0.24321	0.00453	0.002%	73.69%
T24	0.57288	0.83581	69.858%	0.03461	-0.16491	2.720%	72.58%

Table 6
Correlation Coefficients for "Population" of n=50
Subjects from the Holzinger and Swineford (1939) Study

		T5	T6	T12	T13	T23	T24
General Verbal	T5	1.0000	.6440	-.0399	.0762	.3892	.4297
Paragraph Comprehension	T6	.6440	1.0000	.0703	.2587	.5461	.4064
Speeded Dot Counting	T12	-.0399	.0703	1.0000	.3847	.2512	.3705
Speeded Discrimination Caps	T13	.0762	.2587	.3847	1.0000	.3963	.2896
Math Number Series	T23	.3892	.5461	.2512	.3963	1.0000	.5426
Woody-McCall Math	T24	.4297	.4064	.3705	.2896	.5426	1.0000

Table 7
Canonical Function and Structure Coefficients for Sample Data
(n=50)

	Function I			Function II			
Variable/ Coef.	Function	Structure	Squared Structure	Function	Structure	Squared Structure	h^2
T5	0.36603	0.83247	69.301%	-1.25488	-0.55407	30.699%	100.00%
T6	0.72426	0.95999	92.158%	1.08819	0.28002	7.841%	100.00%
Adequacy			80.729%			19.270%	
Redundancy			28.557%			1.714%	
Rc2			35.374%			8.893%	
Redundancy			13.526%			1.780%	
Adequacy			38.238%			20.010%	
T12	-0.31827	0.06100	0.372%	0.43648	0.42443	18.014%	18.39%
T13	0.06941	0.36195	13.101%	0.50749	0.62334	38.855%	51.96%
T23	0.69730	0.90461	81.832%	0.55004	0.35494	12.598%	94.43%
T24	0.47876	0.75926	57.648%	-0.93233	-0.32518	10.574%	68.22%

Table 8
Descriptive Statistics for R_c^2 Across 1,000 Resamplings

Statistic	Function I	Function II
Mean	0.43958	0.13158
SD	0.11744	0.07873
Skewness	0.23703	0.85067
Kurtosis	-0.09252	0.60673

Note. Program CANSTRAP algorithm 1, computation of R_c^2 *independent* of Procrustean rotation of the resampled canonical function matrices, was selected for this analy s.

Table 9
Descriptive Statistics for Function and Structure Coefficients for the Smaller Variable Set Across 1,000 Resamplings

Function Coefficients

*** MEANs		
1	0.4698	-1.1654
2	0.3082	1.0634
*** SDs		
1	0.3643	0.3221
2	0.7101	0.2740
*** SKEWNESSs		
1	1.0118	3.0347
2	-1.2058	-0.4001
*** KURTOSISs		
1	0.5090	12.8658
2	0.0569	0.3770

Structure Coefficients

*** MEANs		
1	0.6552	-0.4587
2	0.6218	0.3074
*** SDs		
1	0.4766	0.3654
2	0.6555	0.2994
*** SKEWNESSs		
1	-1.9942	1.5918
2	-1.8525	0.1646
*** KURTOSISs		
1	2.4850	2.8181
2	1.7368	0.1547

Table 10
Descriptive Statistics for Function and Structure Coefficients
for the Larger Variable Set Across 1,000 Resamplings

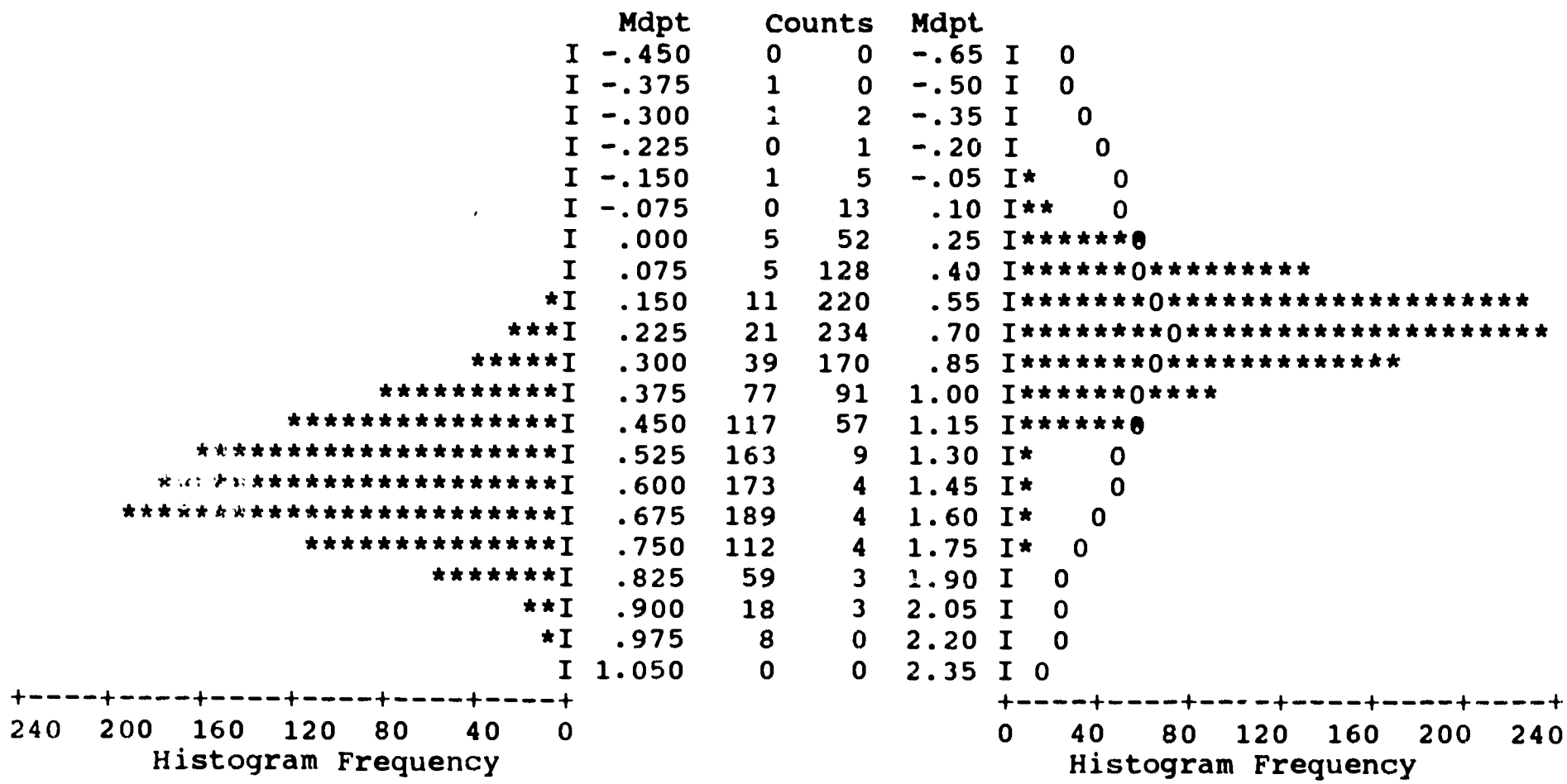
Function Coefficients

*** MEANs		
1	-0.2394	0.3141
2	-0.0147	0.3504
3	0.3953	0.4304
4	0.4196	-0.6641
*** SDs		
1	0.3332	0.4803
2	0.2755	0.3996
3	0.5256	0.3948
4	0.3677	0.4173
*** SKEWNESSs		
1	0.1812	-0.3334
2	0.1025	-0.6003
3	-1.2935	-0.7817
4	-0.6616	1.1365
*** KURTOSISs		
1	-0.4745	-0.3368
2	-0.0576	0.2432
3	0.7095	0.9570
4	0.5870	1.4635

Structure Coefficients

*** MEANs		
1	0.0146	0.2913
2	0.1574	0.4520
3	0.5515	0.3099
4	0.5407	-0.2115
*** SDs		
1	0.2872	0.4174
2	0.3872	0.2901
3	0.6003	0.2962
4	0.4675	0.3532
*** SKEWNESSs		
1	-0.1697	-0.5314
2	-0.6357	-0.7588
3	-1.7525	-0.0483
4	-1.9404	0.4557
*** KURTOSISs		
1	-0.1597	-0.2460
2	-0.3010	0.6731
3	1.5082	-0.1744
4	2.4797	-0.2703

Figure 1
Bootstrap Estimates of r Based on 1,000 Random Resamplings



Note. Each asterisk represents approximately eight cases. The distribution of 1,000 bootstrap estimates of r is presented to the left, while the distribution of the Fisher's Z transformation of these 1,000 estimates is presented to the right. The normal distribution of samples of Z_r, expected given the classical statistical assumptions that sampling error is distributed normally about the estimate, is also presented in the histogram on the right.

APPENDIX A:
Random Sample of n=50 Cases from N=301
(Holzinger & Swineford, 1939, pp. 81-91)

	ID	T5	T6	T12	T13	T23	T24
1	1	40	7	115	229	5	24
2	5	37	8	126	213	4	20
3	9	29	8	93	265	17	18
4	11	33	8	91	157	8	16
5	12	38	6	114	155	5	24
6	13	33	8	103	149	9	25
7	21	41	11	107	177	26	22
8	22	22	5	92	194	23	19
9	30	31	6	117	310	18	27
10	33	29	6	139	215	12	25
11	34	29	5	73	121	9	17
12	36	44	10	87	203	20	22
13	46	28	5	95	100	1	15
14	54	40	11	96	199	18	17
15	66	41	4	123	142	2	15
16	72	54	11	108	227	27	30
17	76	55	13	119	195	6	19
18	78	22	4	115	186	14	21
19	86	26	12	147	207	23	30
20	96	45	12	91	185	23	26
21	98	39	11	119	240	24	23
22	105	48	10	117	152	17	20
23	126	37	8	137	180	10	19
24	136	38	13	139	204	14	18
25	142	45	13	156	252	36	26
26	149	51	10	103	164	16	26
27	156	31	8	112	215	9	18
28	157	44	11	98	139	18	27
29	168	38	9	123	169	10	26
30	203	36	11	86	228	21	18
31	216	56	14	84	171	31	25
32	218	48	11	113	186	24	30
33	219	65	10	104	222	20	20
34	220	49	8	110	161	16	33
35	223	56	13	121	225	23	31
36	224	50	14	115	185	25	32
37	225	25	7	200	236	30	29
38	226	29	8	116	219	6	21
39	233	51	9	113	180	37	27
40	244	29	9	84	137	13	21
41	254	32	5	72	121	4	14
42	277	31	7	97	149	1	20
43	289	48	11	102	224	22	31
44	292	61	10	135	199	28	30
45	297	51	12	110	199	6	26
46	303	36	8	101	179	8	24
47	311	57	13	103	198	19	26
48	323	44	7	140	178	20	29
49	336	48	13	119	195	34	24
50	345	49	9	85	204	17	25