ED 320 965 TM 015 274

AUTHOR Welge-Crow, Patricia A.; And Others

TITLE Looking beyond Statistical Significance: Result Importance and Result Generalizability.

PUB DATE 25 May 90

NOTE 23p.; Paper presented at the Annual Meeting of the American Psychological Society (Dallas, TX, June 9, 1990).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Educational Research; *Effect Size; Estimation (Mathematics); *Generalizability Theory; Heuristics; Mathematical Models; Maximum Likelihood Statistics; *Research Methodology; *Sample Size; *Statistical Significance; *Test Interpretation; Test Results

IDENTIFIERS Empirical Research; Research Replication

ABSTRACT

Three strategies for augmenting the interpretation of significance test results are illustrated. Determining the most suitable indices to use in evaluating empirical results is a matter of considerable debate among researchers. Researchers increasingly recognize that significance tests are very limited in their potential to inform the interpretation of scientific results. The first strategy involves evaluating significance test results in a sample size context. The researcher is encouraged to determine at what smaller sample size a statistically significant fixed effect size would no longer be significant, or conversely, at what larger sample size a non-significant result would become statistically significant. The second strategy would involve interpreting effect size as an index of result importance. The third strategy emphasizes interpretation based on the estimated likelihood that results will replicate. These applications are illustrated via small heuristic data sets to make the discussion more concrete. A 37-item list of references, seven data tables, and an appendix illustrating relevant computer commands are provided. (TJH)

aps.wp 5/25/90

LOOKING BEYOND STATISTICAL SIGNIFICANCE: RESULT IMPORTANCE AND RESULT GENERALIZABILITY

Patricia A. Welge-Crow
LSU Medical Center

Karen B. LeCluyse
University of New Orleans

Bruce Thompson
Texas A&M University

Presented at the annual meeting of the American Psychological Society, Dallas, June 9, 1990.

ABSTRACT

Determining the most suitable indices to use in evaluating empirical results is a matter of considerable debate among researchers (Chow, 1988; Huberty, 1987; Kupfersmid, 1988; Rosnow & Rosenthal, 1989; Thompson, 1989b). Researchers increasingly recognize that significance tests are very limited in their potential to inform the interpretation of scientific results (Carver, 1978). Three strategies for augmenting the interpretation of significance test results are illustrated.

Determining the most suitable indices to use in evaluating empirical results is a matter of considerable debate among researchers (Chow, 1988; Huberty, 1987; Kupfersmid, 1988; Rosnow & Rosenthal, 1989; Thompson, 1989b). Researchers increasingly recognize that significance tests are very limited in their potential to inform the interpretation of scientific results (Carver, 1978). Three strategies for augmenting the interpretation of significance test results are illustrated here.

An Historical Perspective

However, it may be worthwhile to provide an historical perspective on just how far researchers have come in recognizing the potentials and the limits of statistical significance testing. Consider first the position statement of Melton (1962, p. 554) following 12 years of service as editor of the Journal of Experimental Education:

> In editing the Journal there has been a strong reluctance to accept and publish results related to the principal concern of the researcher when those results were [only] significant at the .05 level.... It reflects a belief that it the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounces.

Consider in comparison a statement from one of the several (cf. Kupfersmid, 1988; Meehl, 1978) articles published more recently in prominent journals in psychology:

> It may not be an exaggeration to say that for many PhD students, for whom the .05 alpha has acquired an almost ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation *p* is less than .05.... [But] surely, God loves the .06 nearly as much as the .05 [level]. (Rosnow & Rosenthal, 1989, p. 1277)

Social science has certainly come a long way during the last few years in recognizing the essential limits of significance tests!

Significance as a Test of Sample Size

Even some widely respected authors of prominent textbooks are sometimes not quite sure what role significance tests should play in analysis (Thompson, 1987a, 1988d), 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." A concrete heuristic example may serve to emphasize this point.

Presume that a researcher was working in the Houston school district, and analyzed data involving some of the district's 200,000 students. Perchance the researcher decided to compare the mean IQ scores of 12,000 students located in one zip code with the mean IQ of the 188,000 remaining students residing in other zip codes. Since the *t* distribution approaches the *Z* distribution as sample size approaches infinity, researchers use the *Z* distribution to tests mean differences with large samples. These calculations are reported in Table 1.

INSERT TABLE 1 ABOUT HERE.

The mean IQ (100.15, *SD*=15) of the 12,000 students residing in the zip code of interest differs to a statistically significant degree (Zcalc = 2.12 > Zcrit = 1.96, *p*<.05) from the mean (99.85, *SD*=15) of the remaining 188,000 students. The less thoughtful researcher might suggest to school board members that special schools for gifted students should be erected throughout the zip code of the 12,000 students, since they are "significantly" brighter than their compatriots.

Alternatively, the more thoughtful researcher in such a situation would note that the standardized difference in these two means (.3/15 = 0.02) is trivial. The difference of means (.3 = one-third of one IQ point) is also substantially smaller than one standard error of an IQ measure with a reliability coefficient of

0.92, i.e., SEM = SD*((1-r)**.5) = 4.24. Such a thoughtful researcher would be reticent to extrapolate policy recommendations from every statistically significant result. As Huberty (1987, p. 6) notes, "it would be well to have some idea as to the approximate power (i.e., 1 - beta) one has for some 'important' or 'interesting' alternative hypothesis characterizations, given a particular alpha."

Morrison and Henkel (1970) and Carver (1978) provide historically important and incisive explanations of the limits of significance testing as an aid to interpretation. Although significance is a function of at least seven interrelated features of a study (Schneider & Darcy, 1984), sample size is the primary influence on significance. 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 result in incisive insight that would be lost absent a significance test.

Interpreting Significance Tests in a Sample Size Context

The first strategy for augmenting interpretation of significance tests involves evaluating significance test results in a sample size context. The researcher is encouraged to determine at what smaller sample size a statistically significant fixed effect size would no longer be significant, or conversely, at what larger sample size a nonsignificant result would become statistically significant (Thompson, 1989a).

Table 2 illustrates this application. The table presents significance tests associated with varying sample sizes and large (33.6%) fixed effect sizes. The table can be viewed as presenting results for either a multiple regression analysis involving two predictor variables (in which case the "r sq" effect size would be called the squared multiple correlation coefficient, $\underline{R}^2$) or an analysis of variance involving an omnibus test of differences in three means in a one-way design (in which case the "r sq" effect size would be called the correlation ratio or eta squared).

INSERT TABLE 2 ABOUT HERE.

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

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

Interpreting Effect Size as an Index of Result Importance

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 for an effect can be divided by the SOS total to compute the correlation ratio (also called eta squared). 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.

However, all classical parametric methods are correlational (Knapp, 1978; Thompson, 1988a) and do capitalize on sampling error as part of least squares analyses. Thus, the simpler effect sizes overestimate both the effect size in the full population and the effect size likely to be realized in future studies. 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).

Corrections tend to be larger as either effects sizes or sample sizes become smaller, as illustrated by Thompson (in press). Thus, with a very large effect size or a large sample size, or both, it will matter less which, if any, corrections the researcher applies in estimating effect sizes.

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 <u>third</u> of the variability in the <u>results of studies</u>, [which is distinct from talking about results for only one smaller group of subjects]. (Glass, 1979, p. 13)

<u>Interpreting Results Based on Likelihood of Replication</u>

A third strategy 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 some misconceptions to the contrary, significance tests do not evaluate the probability that results will generalize.

The simplest methods for evaluating replicability partition the sample and then <u>empirically compare results across sample splits</u>. Various sample partitioning methods include conventional cross-validation strategies and also the jackknife methods developed by Tukey and his colleagues (cf. Crask & Perreault, 1977; Daniel, 1989).

The cross-validation methods involve randomly splitting the sample into two subsets, conducting separate analyses, and then empirically comparing the results. Table 3 presents data for a multiple regression example involving two variables ("P" and "R") used to predict the dependent variable ("DV"). The first three subjects were assigned to the first invariance subgroup ("INV"="1"), while the last four subjects were purportedly randomly assigned to the second invariance group. Appendix A presents the SPSS-X commands used to conduct the empirical invariance analysis for these data.

INSERT TABLE 3 ABOUT HERE.

The invariance statistics are produced by the CORRELATIONS procedure. Table 4 presents the invariance results. For this very small data set, the results are not replicable across subsamples. The researcher hopes that invariance coefficients will approach one. Negative values would be very disturbing indeed. But when results appear to be replicable, the researcher can interpret the set of results involving all the subjects with more confidence. The results for all the subjects are always used as the basis for interpretation, since the full sample should theoretically provide the most generalizable results; sample splitting is only performed to evaluate the replicability of the results.

INSERT TABLE 4 ABOUT HERE.

It is very important that result replicability be investigated empirically rather than by subjectively comparing solutions for subsamples. Results can appear very different but actually yield comparable effect sizes. Such cases involve what Cliff (1987, pp. 177-178) refers to as the "sensitivity" of prediction weights.

The most powerful strategy for evaluating result replicability invokes the "bootstrap" methods developed by Efron and his colleagues (cf. Diaconis & Efron, 1983; Efron, 1979; Lunneborg, in press). Conceptually, these 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 (1988c) provides similar software for multivariate applications.

Table 5 presents a small data set that can be used to illustrate "bootstrap" estimation. Table 6 presents descriptive statistics for the data in hand. Table 7 presents "bootstrap" estimates of population correlation coefficients based on the Table 5 data. These estimates were developed using the software available frcır Lunneborg (1987), and were based on 500 samples with replacement.

INSERT TABLES 5, 6 AND 7 ABOUT HERE.

Summary

As Thompson (1989b, p. 4) notes, "significance, importance, and replicability are all important issues in research. 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." The interpretation of empirical results should augment the interpretation of significance tests with (a) interpretation of significance tests in a sample size context; (b) interpretation of effect sizes; and (c) interpretation based on estimated likelihood that results will replicate. These applications were illustrated with small heuristic data sets to make the discussion more concrete.

Refertnces

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

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

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

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

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

Crask, M.R., & Perreault, W.D., Jr. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Daniel, L.G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 305 382)

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.

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

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.

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

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

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

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.

Melton, A. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

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

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.

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

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. (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). 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. (in press). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement.

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
Test of Mean Differences for School District Example

```
Z = ( M1    - M2   ) / (((SD1**2/    n1 ) + (SD2**2/    n2   )) ** .5)
Z = (100.15 - 99.85) / (((15**2 / 12000) + (15**2 / 188000)) ** .5)
  =        0.3       / ((( 225  / 12000) + ( 225  / 188000)) ** .5)
  =        0.3       /  ((    0.01875     +     0.001196     ) ** .5)
  =        0.3       /  (          0.019946808                ** .5)
  =        0.3       /    0.141233170
  =            2.124146887
```

Note. From Thompson (1990), with permission.

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

Source	SOS	r sq	df	MS	F calc	F crit	Decision
SOSexp	**337.2**	**0.336**	**2**	**168.600**	0.253	200.00	Not Rej
SOSunexp	**665.1**		1	665.100			
SOStot	**1002.3**		3	334.100			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	2.535	4.10	Not Rej
SOSunexp	**665.1**		10	66.510			
SOStot	**1002.3**		12	83.525			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	5.070	3.49	Rej
SOSunexp	**665.1**		20	33.255			
SOStot	**1002.3**		22	45.559			
SOSexp	**337.2**	**0.336**	**2**	**168.600**	7.605	3.32	Rej
SOSunexp	**665.1**		30	22.170			
SOStot	**1002.3**		32	31.322			

Note. As sample size increases, tabled "critical *F*" values get smaller. Additionally, as sample size increases, error *df* gets larger, mean square error gets smaller, and thus "calculated *F*" also gets larger. Entries in **bold** remain fixed for the purposes of these analyses. From Thompson (1989b), with permission.

Table 3
Observed and Latent Variables for Small Example Case

P	R	DV	INV	ZP1	ZR1	ZP2	ZR2	YHAT11	YHAT12	YHAT21	YHAT22
1	3	90	1	-1.000	-.132	.	.	.515	-.873	.	.
2	6	49	1	.000	1.060	.	.	-1.152	.304	.	.
3	1	93	1	1.000	-.927	.	.	.637	.570	.	.
4	8	20	2	.	.	-1.162	.669	.	.	-.296	-.779
5	4	3	2	.	.	-.387	-.304	.	.	.474	-.411
6	0	39	2	.	.	.387	-1.276	.	.	1.245	-.042
7	9	63	2	.	.	1.162	.912	.	.	-1.423	1.232

Note. From Thompson (1989b), with permission.

Table 4
Invariance Statistics

	DV	YHAT11	YHAT12	YHAT21
YHAT11	1.0000[a] (n=3)			
YHAT12	-.2842[b] (n=3)	-.2843[c] (n=3)		
YHAT21	-.5182[b] (n=4)	.	.	
YHAT22	.8747[a] (n=4)	.	.	-.5924[c] (n=4)

Note. From Thompson (1989b), with permission.

[a] The multiple correlation coefficient (*R*) for the invariance group.

[b] The "shrunken *R*" for the invariance group.

[c] The invariance coefficient for the invariance group.

Table 5
Data Set for Heuristic Example

n	MILESEC	SYSTOLAV	POND	TOTCHOL	HDLCHOL
1	890(+0.18)	94.0(-1.04)	11.5(-0.91)	180(+0.64)	80.1(+1.17)
2	1097(+1.16)	108.7(+1.42)	12.0(-0.69)	142(-1.56)	51.1(-1.02)
3	1300(+2.12)	97.7(-0.42)	13.1(-0.21)	165(-0.23)	63.3(-0.10)
4	948(+0.45)	90.3(-1.66)	12.6(-0.43)	199(+1.74)	75.7(+0.84)
5	940(+0.41)	100.7(+0.08)	19.3(+2.49)	187(+1.04)	61.0(-0.27)
6	760(-0.44)	104.3(+0.69)	14.7(+0.48)	148(-1.22)	76.0(+0.86)
7	740(-0.53)	95.3(-0.82)	14.2(+0.26)	164(-0.29)	78.5(+1.05)
8	571(-1.33)	97.7(-0.42)	13.6(+0.00)	174(+0.29)	54.3(-0.78)
9	748(-0.50)	102.7(+0.42)	10.9(-1.17)	190(+1.22)	62.2(-0.18)
10	640(-1.01)	96.0(-0.70)	11.4(-0.95)	161(-0.46)	67.4(+0.21)
11	642(-1.00)	107.0(+1.14)	14.6(+0.44)	159(-0.58)	34.8(-2.25)
12	957(+0.49)	108.0(+1.30)	15.2(+0.70)	159(-0.58)	70.8(+0.47)

Note. From Thompson (1990), with permission.

Table 6
Descriptive Statistics and Correlation Coefficients

	MILESEC	SYSTOLAV	POND	TOTCHOL	HDLCHOL	PREDC1	CRITC1
Mean	852.8	100.2	13.6	169.0	64.6	0.0	0.0
SD	211.0	6.0	2.3	17.3	13.2	1.0	1.0
MILESEC		.052	.046	-.047	.140	.063	.048
SYSTOLAV			.244	-.624	-.559	-.981	-.752
POND				.008	-.121	-.084	-.064
TOTCHOL					.243	.637	.830
HDLCHOL						.569	.742
PREDC1							.767

Note. From Thompson (1990), with permission.

Table 7
Bootstrap Estimates of r's for Table 3 Data
Based on 500 Samples with Replacement

Coef.	Table 4 Estimate	Bootstrap Mean	Bootstrap Median	Bootstrap SD
1	0.052	0.0514	0.0417	0.2819
2	0.046	0.0421	0.0603	0.2287
3	-0.047	-0.0233	-0.0489	0.2690
4	0.140	0.1135	0.1551	0.3092
5	0.244	0.2598	0.2428	0.2343
6	-0.624	-0.5878	-0.6196	0.2135
7	-0.559	-0.5430	-0.5737	0.2166
8	0.008	-0.0649	-0.0519	0.3541
9	-0.121	-0.0971	-0.1198	0.2224
10	0.243	0.2189	0.2486	0.2560

Note. From Thompson (1990), with permission.

APPENDIX A: Example SPSS-X Commands for Table 2 Data

```
TITLE 'Demo of Regression Invariance Procedure***'
FILE HANDLE BT/NAME='DEMO7301.DAT'
DATA LIST FILE=BT/P R 1-2 DV 3-4 INV 5
if (inv eq 1)zp1=(p-2.0)/1.0
if (inv eq 1)zr1=(r-3.333)/2.517
if (inv eq 2)zp2=(p-5.5)/1.291
if (inv eq 2)zr2=(r-5.25)/4.113
if (inv eq 1) yhat11=(-.371189*zp1)+(-1.087694*zr1)
if (inv eq 1) yhat12=(.83549*zp1)+(.286434*zr1)
if (inv eq 2) yhat21=(-.371189*zp2)+(-1.087694*zr2)
if (inv eq 2) yhat22=(.83549*zp2)+(.286434*zr2)
variable labels yhat11 'group 1 data using group 1 betas'
                yhat12 'group 1 data using group 2 betas'
                yhat21 'group 2 data using group 1 betas'
                yhat22 'group 2 data using group 2 betas'
print formats zp1 to yhat22 (F8.5)
list variables=p to yhat22
SUBTITLE 'REGRESSION USING ALL DATA'
REGRESSION VARIABLES=P TO DV/DESCRIPTIVES=ALL/DEPENDENT=DV/
  ENTER P R
TEMPORARY
SELECT IF (INV EQ 1)
SUBTITLE 'REGRESSION FOR SUBGROUP #1'
REGRESSION VARIABLES=P TO DV/DESCRIPTIVES=ALL/DEPENDENT=DV/
  ENTER P R
TEMPORARY
SELECT IF (INV EQ 2)
SUBTITLE 'REGRESSION FOR SUBGROUP #2'
REGRESSION VARIABLES=P TO DV/DESCRIPTIVES=ALL/DEPENDENT=DV/
  ENTER P R
subtitle 'check Z calculations'
condescriptive zp1 to yhat22
statistics all
subtitle 'invariance results #######'
correlations variables=dv yhat11 to yhat22/statistics=all
```

Note. The analysis requires two runs. The first run excludes the cards typed in lower case and is conducted to derive the numerical values required for the lower case cards, which are added for the second run.