DOCUMENT RESUME

ED 434 916 TM 030 151

AUTHOR Kieffer, Kevin M.; Thompson, Bruce
TITLE Interpreting Statistical Significance Test Results: A Proposed New "What If" Method.
PUB DATE 1999-11-19
NOTE 25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Point Clear, AL, November 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Research; *Sample Size; *Statistical Significance; *Test Interpretation
IDENTIFIERS *P Values

ABSTRACT

As the 1994 publication manual of the American Psychological Association emphasized, "p" values are affected by sample size. As a result, it can be helpful to interpret the results of statistical significant tests in a sample size context by conducting so-called "what if" analyses. However, these methods can be inaccurate unless "corrected" effect sizes are used. This paper proposes a new method by which "what if" analyses can be conducted using estimated true population effects. Two appendixes contain EXCEL spreadsheet commands for previous "what if" methods and the present method. (Contains 4 tables and 48 references.) (Author/SLD)

msera99.wp1 9/2/99

Interpreting Statistical Significance Test Results:

A Proposed New "What If" Method

Kevin M. Kieffer

Tampa VA Medical Center

Bruce Thompson

Texas A&M University
and
Baylor College of Medicine

Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL, November 19, 1999. The junior author and related reprints may be accessed through Web URL: **"http://acs.tamu.edu/~bbt6147/"**.

Interpreting Statistical Significance Test Results: A Proposed New "What If" Method

Abstract

As the 1994 APA publication manual emphasized, p values are affected by sample size. Thus, it can be helpful to interpret the results of statistical significance tests in a sample size context by conducting so-called "what if" analyses. However, these methods can be inaccurate unless "corrected" effect sizes are employed. The present paper proposes a new method by which "what if" analyses can be conducted using estimated true population effects.

A recent empirical study of journals published since 1950 in education, psychology, medicine, and ecology shows a geometric decade-by-decade growth in the number of articles criticizing statistical significance testing (Anderson, Burnham & Thompson, 1999). The related controversy has even led to a special theme issue of the MSERA journal, Research in the Schools (cf. Daniel, 1998; McLean & Ernest, 1998; Nix & Barnette, 1998). Following scheduled debates at the annual meetings of both the American Psychological Association (APA) and the American Psychological Society, in 1996 the APA Board of Scientific Affairs appointed its Task Force on Statistical Inference (Azar, 1997, 1999; Shea, 1996). The wide-ranging, comprehensive and thoughtful report of the Task Force was published in the summer of 1999 in the American Psychologist (Wilkinson & The Task Force on Statistical Inference, 1999).

In their historical summary dating back to the origins of these tests, Huberty and Pike (in press) provide a thoughtful review of how we got to where we're at as regards statistical tests. Among the recent articles *criticizing* of statistical testing practices, Cohen (1994), Kirk (1996), Rosnow and Rosenthal (1989), Schmidt (1996), and Thompson (1996) have been especially thoughtful. However, these criticisms are certainly not new (see Boring, 1919). Among the classical criticisms, Carver (1978), Meehl (1978), and Rozeboom (1960) in particular have been widely cited.

Among the more thoughtful works *advocating* conventional statistical testing, Cortina and Dunlap (1997), Frick (1996), and especially Abelson (1997), have been most influential. A balanced

and comprehensive treatment of both perspectives is provided by Harlow, Mulaik and Steiger (1997) (for reviews of this book, see Levin, 1998 and Thompson, 1998).

Tenor of More Extreme Conclusions

Two quotations may convey the tenor of some of the conclusions of some more extreme critics of statistical tests. Rozeboom (1997) recently argued that

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

And Tryon (1998) recently lamented,

> [T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial... (p. 796)

Indeed, empirical studies confirm that many researchers do not fully understand the logic of their statistical tests (cf. Mittag,

1999; Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). Misconceptions are taught even in widely-used statistics textbooks (Carver, 1978).

Purpose of the Present Paper

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

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

Indeed, the author guidelines for Measurement and Evaluation in Counseling and Development encourage authors

> ...to assist readers in interpreting statistical significance of their results. For example, results may be indexed to sample size. An author may wish to say, "this correlation coefficient would have still been statistically significant even if sample size had been as small as n = 33," or "this correlation coefficient would have been statistically significant if sample size had been increased to n = 138." (Association for Assessment in Counseling, 1994, p. 143)

Thompson (1989a, 1989b) proposed methods for conducting such "what

if" analyses, which help researchers interpret their results by considering the extent to which sample size (as against effect size) yielded statistical significance.

The purpose of the present paper is to propose a new method for conducting such "what if" analyses to augment the conventional use of statistical significance tests. It will momentarily be noted that the methods previously proposed (Thompson, 1989a, 1989b) have some serious weaknesses, which the new proposed methods address and overcome.

Sample Size Influences

When "nil" null hypotheses (see Cohen, 1994) are used, the null will always be rejected at some sample size. There are infinitely many possible sample effects (Kirk, 1996), and therefore the probability of obtaining an exactly zero sample effect, as specified by a "nil" null, is infinitely small. That is, because probability equals the given occurrence divided by the total number of occurrences, given that the denominator is infinite, the probability is infinitely small. Therefore, given a "nil" null, and a non-zero sample effect, the null hypothesis will <u>always</u> be rejected at some sample size!

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

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether

there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. (Thompson, 1992, p. 436)

Certainly this dynamic is well known, even though its influence is just as widely underestimated. Table 1 illustrates how sample size impacts $p_{CALCULATED}$ values for a hypothetical one-way four-level ANOVA study (or, alternatively, a regression study involving three predictor variables). These results involve an eta^2 (η^2) value, 13.8%, that Cohen (1988, pp. 26-27) characterized as "large" as regards result typicality. [In an ANOVA eta^2 is the percentage of variance in the dependent variable that can be predicted with knowledge of the participants' memberships in the study's groups or design cells.]

INSERT TABLE 1 ABOUT HERE.

More to the point, statistical significance testing with "nil" null hypotheses is arguably *irrelevant either when (a) sample size is very large or (b) effect size is very large*. Table 2 presents illustrative results here for the Pearson product-moment correlation coefficient (including bivariate reliability or validity coefficients). The first three effect sizes were those characterized by Cohen (1988, pp. 24-27) as "low," "medium," and "large," as regards result typicality.

INSERT TABLE 2 ABOUT HERE.

More than 60 years ago, Berkson (1938) wrote an article titled, "Some difficulties of interpretation encountered in the

application of the chi-square test." He noted that when working with data from roughly 200,000 people,

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

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

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

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

Thus, Table 2 also illustrates that reliability or validity

coefficients of this magnitude will be statistically significant even with n's as small as four or five people.

The Old "What If" Method

Table 3 illustrates the use of the "what if" method described by Thompson (1989a, 1989b). In this method certain values are taken as fixed (presented in **bold** in the table). The table presumes a regression study involving two predictor variables ($df_{EXPLAINED}$ = 2) and n=40. The first three rows in the table present hypothetical results from this study.

Then the sample size is varied, assuming a fixed effect size (e.g., R^2, eta^2). [The tabled results here could equally well be viewed as a three-level one-way ANOVA problem ($df_{EXPLAINED}$ = k-1 = 2), because R^2 = $SOS_{EXPLAINED}$ / SOS_{TOTAL}, but also eta^2 = $SOS_{EXPLAINED}$ / SOS_{TOTAL}.] In the Table 3 example, given the design and a fixed variance-accounted-for effect size of 10.0%, the effect size becomes statistically significant (α=.05) when n goes from 59 to 60. These analyses can be easily conducted using a microcomputer spreadsheet such as Excel; Appendix A presents the commands.

INSERT TABLE 3 ABOUT HERE.

A New Proposed Method

"Corrected" vs "Uncorrected" Effect Sizes

"Classical" statistical methods (e.g., ANOVA, regression) use the statistical theory called "ordinary least squares." This theory optimizes the fit of the synthetic/latent variables (e.g., $\hat{Y}$) to the observed/measured outcome/response variables (e.g., Y) in the **sample** data, and capitalizes on all the variance present in the

observed sample scores, including the "sampling error variance" that it is idiosyncratic to the particular sample. Because sampling error variance is unique to a given sample (i.e., each sample has its own sampling error variance), "uncorrected" variance-accounted-for effect sizes (e.g., R^2, eta^2) somewhat overestimate the effects that would be replicated in either (a) the population or (b) a future sample.

However, statistical theory can be invoked to estimate the extent of overestimation (i.e., positive bias) in the sample variance-accounted-for effect size estimate. [Note that "corrected" estimates are always less than or equal to "uncorrected" values.] The difference between the "uncorrected" sample (e.g., R^2, eta^2) and "corrected" population (e.g., "adjusted R^2", Hays' omega^2 (ω^2)) variance-accounted-for effect sizes is called "shrinkage." That is, these "corrected" effect size estimates are estimates of the effect size in the **population**.

For example, for regression the "corrected" effect size "adjusted $\underline{R}^2$" is automatically provided by most statistical packages, even without being requested. This correction is due to Ezekiel (1930), although the formula is often incorrectly attributed to Wherry (Kromrey & Hines, 1996):

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

where $\underline{n}$ is the sample size and $\underline{v}$ is the number of predictor variables. For example, if $\underline{n}$ = 60, and there are six predictor variables in a regression analysis for which R^2 = 50.0%, "adjusted R^2"

$$= 1- [(n - 1) / (n - v - 1)] * (1 - R^2)$$

$$
\begin{aligned}
&= 1- [(60 - 1) / (60 - 6 - 1)] * (1 - .5) \\
&= 1- [(\ 59\ ) / (60 - 6 - 1)] * (1 - .5) \\
&= 1- [(\ 59\ ) / (\ \ 53\ \ )] * (1 - .5) \\
&= 1- [(\ 59\ ) / (\ \ 53\ \ )] * (\ \ .5\ \ ) \\
&= 1- [\ \ \ 1.113\ \ \ \ \ \ ] * (\ \ .5\ \ ) \\
&= 1- \qquad 0.556 \\
&= .44339 = 44.3\%.
\end{aligned}
$$

The "adjusted R^2" formula can also be equivalently expressed as:

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

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

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

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

Problem with the Previous (Thompson, 1989a, 1989b) Method

The previously proposed "what if" methods (Thompson, 1989a, 1989b), as illustrated in Table 3, are problematic in that they are conducted in the metric of the sample (i.e., an "uncorrected" effect size) rather than in the metric of the population from which all samples are ostensibly drawn (i.e., a "corrected" effect size). Put differently, *the previous methods do not take into account that the amount of sampling error (and therefore the positive bias in the "uncorrected" effect size) will change as sample size itself changes.*

Proposed Alternative "What If" Method

The alternative "what if" analytic method proposed here invokes the "corrected" estimate of the *population* effect size as the metric for exploring sample size influences. This analysis is illustrated in Table 4 using the same research design and results presumed in the first three rows of Table 3. Appendix B presents

the spreadsheet commands that readily implement the proposed analyses.

INSERT TABLE 4 ABOUT HERE.

In these analyses "adjusted R^2" (or omega^2) is taken as fixed, while R^2 (or eta^2) changes along with sample size variations. R^2 can be solved for, given "adjusted R^2," by algebraically rearranging the Ezekiel (1930) correction formula. If, for example, "adjusted R^2" = 44.3%, $\underline{n}$ = 60, and there are six regression predictor variables, ("uncorrected") R^2

```
=  1-[-1*[(Adj R² - 1) / [( n - 1) / ( n  - v - 1)] ]]
=  1-[-1*[(.4433  - 1) / [(60 - 1) / ( 60 - 6 - 1)] ]]
=  1-[-1*[(.4433  - 1) / [(60 - 1) / (     53    )] ]]
=  1-[-1*[(.4433  - 1) / [(  59  ) / (     53    )] ]]
=  1-[-1*[(  -.556   ) / [(  59  ) / (     53    )] ]]
=  1-[-1*[(  -.556   ) / [          1.113          ] ]]
=  1-[-1*[                  -.5                     ]]
=  1-[                       .5                      ]
=  .5 = 50.0%.
```

Discussion

The illustrative results presented in Tables 3 and 4 make clear how different can be the results from the two "what if" analytic strategies. In the illustrations, a researcher obtains a squared multiple correlation coefficient (R^2) of 10.0% in a hypothetical study involving two predictor variables and 40 participants.

Using the classical "what if" analyses proposed by Thompson (1989a, 1989b), this unadjusted effect size for this design becomes statistically significant (α=.05) when $\underline{n}$ goes from 59 to 60, as reported in Table 3. However, for the same design, using what is suggested here is the more accurate new proposed "what if" analytic

strategy, the results become statistically significant (α=.05) when n goes from 121 to 122, as reported in Table 4.

Applicability with Results that Were Originally Significant

The previous discussion has involved a hypothetical research scenario in which the original results were not statistically significant. However, it is emphasized that the proposed strategy is not limited to use with results that were originally non-significant.

The proposed strategy can be just as useful with results that were originally statistically significant, just as the previous "what if" methods could also be employed with just such results (Thompson, 1989a, 1989b). As the author guidelines for Measurement and Evaluation in Counseling and Development note, for example,

> An author may wish to say, "this [statistically significant] correlation coefficient would have still been statistically significant even if sample size had been as small as n = 33"... (Association for Assessment in Counseling, 1994, p. 143)

Benefits of "What If" Analyses

Use of these "what if" methods may prevent authors with large sample sizes from overinterpreting their small effects, once they see that the small effects would no longer have been statistically significant with even only a slightly smaller sample size. Conversely, researchers with large effects will be even more confident in interpreting their results if they note that their observed effects would still have been statistically significant even if they had had an appreciably smaller sample size.

Example Applications for Recently Published Articles

Small sample size, small effect size. Lightsey and Christopher (1997) conducted a study of what variables predicted variability in the depression of 60 participants. For example, they reported that thinking positive thoughts had an r^2 effect size of 5%. The result was not statistically significant (α=.05).

However, a "what if" analysis of the results, given the design, and using the proposed methods and the appended spreadsheet commands, indicated that the result would have been statistically significant if only 2 more people had participated in the study, assuming that the effect size would then be roughly the same. The proposed "what if" analysis suggests that the result may still be noteworthy, and remind us that "surely, God loves the .06 nearly as much as the .05" level of statistical significance (Rosnow & Rosenthal, 1989, p. 1277).

Large Effect Size. In a study of 195 womens' perceptions of counselors with one of three different orientations as regards feminism, Hackett, Enns and Zetzer reported a three-level one-way ANOVA result for which the $F_{CALCULATED}$ value was 121.74 (df = 2, 192). By employing the proposed new methods and appended spreadsheet commands, it was determined that the eta^2 value for this result was 55.9%, while the adjusted effect size was 55.2%.

This effect size is several times larger than the effect size that Cohen (1988) characterized as "large." The "what if" analysis of the result makes clear just how large the effect was. The results indicate that this effect size would have remained statistically significant with an n as small as 7 people!

Large Sample Size. For the first major hypothesis that she tested, Voelkl (1995) had a sample size of 13,098 participants for a multiple regression problem involving four predictor variables. She reported that, "The overall test of association between school warmth and four student achievement measures was statistically significant, [sic] $F(4, 13093) = 23.83$, $p < .0001$" (p. 133). For these results we can use a computer spreadsheet to determine that R^2 was 0.72% and the actual $p_{CALCULATED}$ value was 1.2E-19 (i.e., 19 zeroes to the right of the decimal, followed by "12"). The author's discussion of this result was quite succinct, and was offered without any reference to the influence of sample size: "Student perceptions of school warmth were significantly related to academic achievement" (p. 136).

By employing the proposed new methods using the Appendix B spreadsheet commands, we find that here the sample results for a population "adjusted R^2" (i.e., 0.68%) remains statistically significant until the n drops from 657 to 656. The result clearly indicates that, although the original sample size was huge, still a very large sample remains necessary for this very small effect size to remain statistically significant (α=.05). This result from the proposed "what if" analysis reinforces the notion that p values *cannot* be used as reasonable indices of effect, and that the results for this hypothesis were not noteworthy even though they were statistically significant.

Caveat

The proposed analyses and related methods (see Morse, 1999) do not constitute either magic or a panacea. But the proposed methods

may help researchers to see how their sample size may have impacted their calculated p values.

It is also very important to emphasize that the analyses do not have to presume that an exact effect size would be replicated in a future study with more (or fewer) participants. Indeed, in addition to conducting the analyses illustrated here, it may be very useful to conduct the same analyses with both somewhat larger and somewhat smaller effect sizes, so as to model sample size impacts for a given design across a reasonable range of effect size outcomes!

As suggested elsewhere,

> In any case, the purpose of this approach is not to identify the exact results that would occur with a different sample size, assuming exactly the same effect size. Rather, the approach focuses on establishing a general ballpark for interpreting statistical significance tests in a sample size context. (Thompson, 1993, p. 368)

Thus, "what if" analysis, just like any other analyses, should not be overinterpreted, and must be employed reasonably and reflectively.

References

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

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

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

Association for Assessment in Counseling. (1994). Guidelines for authors. Measurement and Evaluation in Counseling and Development, 27, 341.

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

Azar, B. (1999). APA statistics task force prepares to release recommendations for public comment. The APA Monitor, 30(5), 9.

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

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

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

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

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

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

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

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools, 5(2), 23-32.

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

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

Hackett, G., Enns, C.Z., & Zetzer, H.A. (1992). Reactions of women to nonsexist and feminist counseling: Effects of counselor orientation and mode of information delivery. Journal of Counseling and Development, 39, 321-330.

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

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

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

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

Kirk, R. (1996). Practical significance: A concept whose time has

come. Educational and Psychological Measurement, 56, 746-759.

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

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

Lightsey, O.W., Jr., & Christopher, J.C. (1997). Stress buffers and dysphoria in a non-western population. Journal of Counseling and Development, 75, 451-459.

McLean, J.E., & Ernest, J.M. (1998). The role of statistical significance testing in educational research. Research in the Schools, 5(2), 15-22.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

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

Morse, D.T. (1999). MINSIZE2: A computer program for determining effect size and minimum sample size for statistical significance for univariate, multivariate, and nonparametric tests. Educational and Psychological Measurement, 59, 518-531.

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

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: ban

or abandon. A review of null hypothesis significance testing. Research in the Schools, 5(2), 3-14.

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

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

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

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

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

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

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

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22(2), 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(1), 2-5.

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

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

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

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

Tryon, W.W. (1998). The inscrutable null hypothesis. American Psychologist, 53, 796.

Voelkl, K.E. (1995). School warmth, student participation, and achievement. Journal of Experimental Education, 63, 127-138.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: **http://www.apa.org/journals/amp/amp548594.html**]

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

Table 1
Illustrative Variations in p for a **Fixed** Effect Size
Across Various Sample Sizes

	Statistic			
n	eta^2	$omega^2$	F_{calc}	p_{calc}
6	13.8%	-80.7%	0.11	.9487
12	13.8%	-13.6%	0.48	.7040
18	13.8%	-3.3%	0.80	.5127
24	13.8%	1.4%	1.12	.3632
30	13.8%	4.1%	1.44	.2528
36	13.8%	5.8%	1.76	.1738
42	13.8%	7.0%	2.08	.1184
48	13.8%	7.9%	2.40	.0801
54	13.8%	8.6%	2.72	.0539
60	13.8%	9.1%	3.04	.0361
66	13.8%	9.6%	3.36	.0241
72	13.8%	9.9%	3.68	.0160
78	13.8%	10.2%	4.00	.0106
84	13.8%	10.5%	4.32	.0070
90	13.8%	10.7%	4.64	.0047
96	13.8%	10.9%	4.96	.0031
102	13.8%	11.1%	5.28	.0020

Table 2
Illustrations of How Statistically Significant Results Occur
Either with Large n's and Small Effects
or Large Effects and Small n's

	Statistic			
n	r^2	$omega^2$	F_{calc}	p_{calc}
385	1.0%	0.7%	3.87	**.0499**
65	5.9%	4.4%	4.01	**.0494**
28	13.8%	10.3%	4.32	**.0472**
4	70.0%	54.5%	7.00	.0773
5	70.0%	58.1%	9.33	**.0378**
3	80.0%	63.6%	8.00	.1056
4	80.0%	68.8%	12.00	**.0405**
3	90.0%	81.0%	18.00	.0513
4	90.0%	83.9%	27.00	**.0138**

Note. Sample sizes (n) and p values less than .05 are presented in bold. The first three effect sizes were those characterized by Cohen (1988, pp. 24-27) as "low," "medium," and "large," as regards result typicality.

Table 3
Illustration of Previously Proposed (Thompson, 1989a, 1989b) "What If" Analyses

Source	SOS	df	MS	F_{calc}	p_{calc}	R^2	Adj R^2
n = 40							
Exp	**10.000**	**2**	**5.000**	2.056	.142392	**10.00%**	2.50%
Unexp	**90.000**	37	2.432				
Total	**100.000**	39	2.564				
n = 59							
Exp	**10.000**	**2**	**5.000**	3.111	.052335	**10.00%**	5.09%
Unexp	**90.000**	56	1.607				
Total	**100.000**	58	1.724				
n = 60							
Exp	**10.000**	**2**	**5.000**	3.167	.049649	**10.00%**	5.18%
Unexp	**90.000**	57	1.579				
Total	**100.000**	59	1.695				

Note. Fixed values in the "what if" analysis are presented in **bold**.

Table 4
Illustration of New Proposed "What If" Methods Using Fixed **"Corrected"** Effect Size

Source	SOS	df	MS	F_{calc}	p_{calc}	R^2	Adj R^2
n = 40							
Exp	10.000	**2**	5.000	2.056	.142392	10.00%	**2.50%**
Unexp	90.000	37	2.432				
Total	**100.000**	39	2.564				
n = 60							
Exp	7.458	**2**	3.729	2.297	.109827	7.46%	**2.50%**
Unexp	92.542	57	1.624				
Total	**100.000**	59	1.695				
n = 80							
Exp	6.203	**2**	3.101	2.546	.084988	6.20%	**2.50%**
Unexp	93.797	77	1.218				
Total	**100.000**	79	1.266				
n = 121							
Exp	4.937	**2**	2.469	3.064	.050413	4.94%	**2.50%**
Unexp	95.063	118	.806				
Total	**100.000**	120	.833				
n = 122							
Exp	4.917	**2**	2.459	3.077	.049777	4.92%	**2.50%**
Unexp	95.083	119	.799				
Total	**100.000**	121	.826				

Appendix A
Excel Spreadsheet Commands
for Previous "What If" Method (Table 3)

Input:

	A	B	C	D	E	F	G	H
1	Source	SOS	df	MS	Fcalc	pcalc	R^2	Adj R^2
2	Exp	+b4*g2	2	+b2/c2	+d2/d3	**Note F1**	10.00%	**Note H1**
3	Unexp	+b4-b2	+c4-c2	+b3/c3				
4	Total	100.000	+c5-1	+b4/c4				
5		n =	40					

Note. Cell **F1** = =fdist(e2,c2,c3)
Cell **H1** = =1-(((+c4)/(+c4-(c2+1)))*(1-g2))

Output:

	A	B	C	D	E	F	G	H
1	Source	SOS	df	MS	Fcalc	pcalc	R^2	Adj R^2
2	Exp	10.000	2	5.000	2.056	.142392	10.00%	2.50%
3	Unexp	90.000	37	2.432				
4	Total	100.000	39	2.564				
5		n =	40					

Appendix B
Excel Spreadsheet Commands
for Proposed "What If" Method (Table 4)

Input:

	A	B	C	D	E	F	G	H
1	Source	SOS	df	MS	Fcalc	pcalc	R^2	Adj R^2
2	Exp	+g2*b4	2	+b2/c2	+d2/d3	**Note F1**	**Note G1**	2.50%
3	Unexp	+b4-b2	+c4-c2	+b3/c3				
4	Total	100.000	+c5-1	+b4/c4				
5		n =	122					

Note. Cell **F1** = =fdist(e2,c2,c3)
Cell **G1** = =1-(-1*((h2-1)/((c4)/(c4-(c2+1)))))

Output:

	A	B	C	D	E	F	G	H
1	Source	SOS	df	MS	Fcalc	pcalc	R^2	Adj R^2
2	Exp	4.917	2	2.459	3.077	.049777	4.92%	2.50%
3	Unexp	95.083	119	.799				
4	Total	100.000	121	.826				
5		n =	122					

TM030151

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

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: INTERPRETING STATISTICAL SIGNIFICANCE TEST RESULTS: A PROPOSED NEW "WHAT IF" METHOD	
Author(s): KEVIN M. KIEFER and BRUCE THOMPSON	
Corporate Source:	Publication Date: 11/19/99

II. REPRODUCTION RELEASE:

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

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

☒ ⬅ Sample sticker to be affixed to document

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

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRUCE THOMPSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ➡ ☐

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

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

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

Signature:	Position: PROFESSOR
Printed Name: BRUCE THOMPSON	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1335
	Date: 9/2/99

OVER