DOCUMENT RESUME

ED 427 090		TM 029 473

AUTHOR	Hetrick, Sam
TITLE	A Primer on Effect Sizes: What They Are and How To Compute Them.
PUB DATE	1999-01-00
NOTE	13p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computation; *Effect Size; Meta Analysis; *Research Methodology; Statistical Significance

ABSTRACT

Magnitude of effect (ME) statistics are an important alternative to statistical significance. Why methodologists encourage the use of ME indices as interpretation aids is explained, and different types of ME statistics are discussed. The basic concepts underlying effect size measures are reviewed, and how to compute them from published reports even when results are incompletely reported is explained. Effect size measures are increasingly important, especially since the American Psychological Association publication manual explicitly suggests that they be reported. (Contains 25 references.) (SLD)

A Primer on Effect Sizes: What They Are and How to Compute Them

Sam Hetrick

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

Abstract

The paper reviews the basic concepts underlying effect size measures, and how to compute them from published reports even when results are incompletely reported. Such measures are increasingly important, especially with the APA publication manual (1994, p. 18) explicitly encouraging that effect sizes always are reported.

A Primer on Effect Sizes: What They Are and How to Compute Them

Statistical significance testing is a prominent feature of data analytic traditions in the social sciences. For many years, methodologists have debated what statistical significance testing means and how it should be used in the interpretation of substantive results (e.g., Carver, 1978; Greenwald, 1975; Hays, 1963; Meehl, 1978; Morrison & Henkel, 1970; Thompson, 1989). The authors of a series of articles appearing in recent editions of the American Psychologist continue the discussion of statistical significance testing and common, persistent misconceptions associated with this tradition (e.g., Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989).

Especially noteworthy are recent articles by Cohen (1994), Kirk (1996), Schmidt (1996), and Thompson (1996). Also, as noted in the August 16, 1996 issue of the Chronicle of Higher Education (pp. A12 and A17), APA has now created a Task Force on Statistical Inference which will consider various proposals, including banning statistical significance testing in APA journals.

The purpose of the present paper is to discuss the use of magnitude of effect (ME) statistics as one alternative for statistical significance. I explain why methodologists encourage the use of ME indices as interpretation aids and discuss different types of ME statistics. Also discussed are correction formulas developed to attenuate statistical bias in ME estimates, and the effect of these formulas on different sample and effect sizes are illustrated (cf. Snyder & Lawson, 1993).

Statistical Significance versus Importance

Use of an instructional method that increases the performance of an experimental group on a dependent measure by 5 points over a control group will result in statistically significant findings, if sample size is large enough. Whether or not such a 5 point difference (i.e., magnitude of effect) between the groups is meaningful from an instructional standpoint depends on many factors other than the statistically significant p value.

It is critical that researchers recognize that a small p value does not necessarily imply that the strength of the relation between the independent and dependent variables in a particular study is large (Rosnow & Rosenthal, 1989). Systematic examination of the magnitude of the effect can assist the researcher in determining how much sample size is influencing results. Although achieving statistical significance is a function of at least seven interrupted study features (Schneider & Darcy, 1984), sample size is the primary influence on whether or not results will be statistically significant. As Craig, Eison, and Metze (1976) noted, "Given a large enough sample size, a significant result may be identified when there is very little association between the independent and dependent variables" (p. 280). As Hays (1963) argued:

> [T]he occurrence of a significant result says nothing at all about the strength of association between treatment and scores. A significant result leads to the inference that some association exists, but in no sense does this mean that an important degree of association necessarily exists. Conversely, evidence of a strong statistical

> association can occur in data even when the results are not significant. The game of inferring the true degree of statistical associations has a joker: [T]his is the sample size. The time has come to define the notion of the strength of the statistical association more sharply, and to link this idea with that of the true difference between population means (p. 324)

Fallacies of Statistical Significance Testing

For almost 70 years, social scientists have shared a seeming obsession with null hypothesis significance testing. Although the usefulness of this method has been refuted for nearly as many years, it still remains the primary method used to interpret data (Kirk, 1996). Simply because results are deemed to be statistically significant, that does not mean that they are intrinsically interesting. Obtaining statistically significant results does not mean that the results are replicable or have any clinical or practical significance.

Considering how the null hypothesis is always false (Cohen, 1994; Thompson, 1996), the use of null hypothesis significance testing appears moot. If nonsignificant p values are assessed, all that that means is that the sample size was not large enough to obtain statistically significant results. Likewise, if statistically significant results are in fact obtained, that only means that we know only the direction of the difference between the control and treatment groups while remaining ignorant of the extent of them (Kirk, 1996).

Reforms have been posited that can be used as other ways to interpret data (Thompson, 1989, 1996). These include the jackknife, bootstrap, and cross-validation methods. The jackknife is a process in which different subjects are dropped from analysis to determine how consistent the results are across different scenarios of omitted subjects. In the bootstrap method, the data is recopied multiple times into a megafile. Then different samples are drawn from the megafile to determine the effect of sampling. Cross-validation methods are used by randomly dividing the subjects into two subsets and then analyzing the two subgroups separately.

The Alternative Method of Using Cohen's d

Although p has been the primary statistic used to interpret data, other more useful techniques have been devised. In 1969, Cohen introduced the concept of d and it has remained one of the most noteworthy alternatives to p that has been utilized in social sciences. This method does not require any more information that does the use of p test, but proves to be much more useful.

One of the flaws of null hypothesis significance testing is the black-or-white, all-or-nothing logic that it uses. Either the researcher rejects or fails to reject the null hypothesis. Considering how the usefulness of a particular treatment is not always so black or white, the extent of effectiveness should be considered. Cohen set out guidelines for determining the magnitude of d. He divided the range into small, medium, and large effects (Kirk, 1996). A medium d of .5 is considered to noticeable while a small one of .2 is deemed nontrivial. A value of .8 was set aside for a large effect size because it was the same distance from the medium value as the small amount of .2 is. Although these values are useful in determining the value of d, Kirk (1996) describes how social scientists should not unquestionably obey these values in a rigid manner. Subjective discretion should be exercised when considering the practical significance of these values.

Another Alternative: Variance-accounted-for Effect Sizes

Standardized differences, such as Cohen's d, can be readily computed for experiments involving two groups where the researcher is focusing on means. However, in non-experiments, or studies with more than two groups, or where statistics other than means are of interest, variance-accounted-for effect sizes (e.g., eta^2, $omega^2$, R^2, adjusted R^2) analogous to r^2 can always be computed (see Snyder & Lawson, 1993). Indeed, these effect sizes can be computed in any analysis, because all analyses are correlational (cf. Fan, 1996, 1997; Knapp, 1978; Thompson, 1984, 1991, in press).

Shortcomings of Effect Size Estimates

Effect size estimates are only as useful as the researcher who interprets them. Only through the proper interpretation of Cohen's d and other effect sizes can useful insight be obtained. Magnitude-of-effect statistics, like any other form of statistics, are context dependent. Snyder and Lawson (1993) posit that despite Cohen's differentiation of small, medium, and large effect sizes, "the judgment regarding the clinical significance of an ME ultimately rests with the researcher's personal value system, the research questions posed, societal concerns, and the design of a particular study."

Although interpretation of p apparently requires researchers to rigidly pay homage to numbers that have been arbitrarily set, such as .05 and .01, interpretation of effect sizes does not share similar fixations. Cohen's values of .02, .05, and .08 are merely suggestions and should not be viewed a magic numbers. As Snyder and Lawson (1993) argued, "Setting arbitrary guidelines against which to evaluate the size of a *particular* ME discounts the context dependency of the investigative process" (p. 347).

<u>Summary</u>

The traditional use of null hypothesis statistical significance testing obviously has many inherent flaws. Primarily, it does not serve as an indicator or whether or not any practical or clinical significance can be derived from the data. Other methods, such as the use of the jackknife, the bootstrap, and cross-validation methods provide possible ways to reform this traditional yet possibly misleading form of data analysis.

Misuses of statistical significance tests remain endemic notwithstanding withering criticisms of these abuses (cf. Cohen, 1994; Kirk, 1996; Rosnow & Rosenthal, 1989; Schmidt, 1996; Thompson, 1996). Thus, a few have 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. . . . (Rozeboom, 1997, p. 335)

Similarly, Tyron (1998) recently noted,

> [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)

The most promising alternative to statistical significance lies with the use of effect sizes. Snyder and Lawson (1993) provide an excellent review of these methods. The most common form of effect size interpretation is the use of Cohen's $\underline{d}$ in which effects can be determined to be either small, medium, or large. Nonetheless, the use of effect sizes, like any other form of statistics can be misleading is not interpreted properly.

References

Carver, R. (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, 1304-1312.

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

Craig, J.R., Rison, 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.

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

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

Greenwald, A. (1975). Consequences of prejudices against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W.L. (1963). Statistics for psychologists. New York: Holt, Rinehart, and Winston.

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

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

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

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.

O'Grady, K.E. (1982). Measures of explained variance: Cautions and limitations. Psychological Bulletin, 92, 766-777.

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

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

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

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

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

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

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

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

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-5.

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

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

TM029473

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

REPRODUCTION RELEASE
(Specific Document)

ERIC

I. DOCUMENT IDENTIFICATION:

Title: A PRIMER ON EFFECT SIZES: WHAT THEY ARE AND HOW TO COMPUTE THEM	
Author(s): SAM HETRICK	
Corporate Source:	Publication Date: 1/99

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.

[X] 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
SAM HETRICK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document [ ]

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

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: X Sam Hetrick	Position: RES ASSOCIATE
Printed Name: SAM HETRICK	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 1/20/99

OVER