DOCUMENT RESUME

ED 449 212 TM 032 357

AUTHOR Cromwell, Susan
TITLE An Introductory Summary of Various Effect Size Choices.
PUB DATE 2001-02-01
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Research Methodology

ABSTRACT

This paper provides a tutorial summary of some of the many effect size choices so that members of the Southwest Educational Research Association would be better able to follow the recommendations of the American Psychological Association (APA) publication manual, the APA Task Force on Statistical Inference, and the publication requirements of some journals. Effect sizes can be classified into two general families: standard differences and variance-accounted-for measures of strength of association. Standardized differences are defined as the standardized difference between two groups. The variance-accounted-for measure of strength of association is defined as the variance-accounted-for squared correlation between the independent and dependent variables. Within these families, several different choices of effect size are available. (Contains 22 references.) (Author/SLD)

Running head: SUMMARY OF VARIOUS EFFECT SIZES

An Introductory Summary of Various Effect Size Choices

Susan Cromwell

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1, 2001.

Abstract

The purpose of the present paper is to provide a tutorial summary of some of the many effect size choices, so that SERA members will be better able to follow the recommendations of the APA publication manual, the APA Task Force on Statistical Inference, and the publication requirements of some journals. Effect size can be classified into two general families; standard differences and variance-accounted-for measures of strength of association. Within both families, several different choices of effect sizes are available (Snyder & Lawson, 1993).

An Introductory Summary of Various Effect Size Choices

Over the years, statistical significance has been the prominent feature of data analyses in the field of education and other social sciences. However, statistical significance tests do not always (if ever) aid the researcher in determining whether results are of practical significance. Thus, the frequencies of publications of criticisms of statistical testing have grown exponentially decade by decade across diverse disciplines (Anderson, Burnham & Thompson, 2000).

Kirk (1996) pointed out three main areas of criticism concerning classical null hypothesis significance testing. First, statistical significance tests do not tell the researcher what they want to know. The researcher wants to know the probability of the null hypothesis being true in the population, but testing the significance of the null hypothesis tells the researcher the probability of obtaining sample data that supports the null hypothesis if the null hypothesis is assumed true in the population. The second criticism is that statistical significance testing is a trivial exercise because there will always be some degree of difference between the two variables; therefore, statistical significance can always be met depending on the power of the research study (Thompson & Keiffer, 2000). The important part that is often overlooked is whether or not the effect is useful or large enough to make a practical difference, regardless of the level of statistical significance. This led to researchers following the rules of null hypothesis statistical testing to such a narrow degree that researchers focused on controlling the Type I error that cannot occur, because essentially

all null hypotheses are false, while causing the Type II errors that can occur to exceed acceptable levels. Third, by setting a predetermined level of statistical significance, researchers can obtain statistical significance simply by manipulating the sample size. The higher the sample size, the more likely the researcher will find statistical significance. This dynamic can create a tautology.

Due to major criticisms such as stated above, methodologists suggested researchers use magnitude-of-effect estimates in result interpretation to highlight the distinction between statistical and practical significance. Practical significance is an alternative to statistical significance when interpreting the outcome of research or studying theory development. With statistical significance when the null hypothesis is false, the researcher is simply unable to specify the direction of the difference between A and B. Now, with a rejection of the null hypothesis, the researcher can be almost certain of the direction of the difference. However, being almost certain can be considered unscientific, and it seems more like a gamble on where the difference may occur. And we also care (very much) about how big the effect is.

Take smoking for example. Taking a gamble on whether smoking causes cancer seems unethical, yet it is what happens when the only focus is on statistical significance because ignoring the size of difference is exactly where the importance lies. What can give more insightful results, and aid in possible applications of research through a more scientific and ethical manner? Practical significance, which involves finding the size of the difference and the

error associated with the estimated difference (cf. Kirk, 1996). This can be accomplished by using a magnitude-of-effect estimate. A magnitude-of-effect estimate (i.e., effect size) tells to what degree the dependent variable can be controlled, predicted, or explained by the independent variables(s) (Olejnik & Algina, 2000; Snyder & Lawson, 1993). Thompson (in press) provides a comprehensive review of modern effect size choices.

Various types of effect size exist, and there are two reasons learning about this statistical area is so vital. One, the researcher needs to be informed of alternative statistical measures that more accurately interprets and reports differences within the results, other than null hypothesis statistical testing characterize results. Also, an understanding must be obtained about the difference between the terms statistical significance and importance. Unfortunately, these words are often used synonymously. Effect size statistics assist the researcher in the clarification of whether statistically significant findings may be practical, or important, when compared to the actual research topic (Snyder & Lawson, 1993). Second, it is vital to researchers to be better prepared to follow the new guidelines concerning effect size set by APA. Examples include the recommendations of the APA publication manual (Kirk, 2001; Shibley Hyde, 2001; Vacha-Haase, 2001), the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999), and the publication requirements of some journals (Kieffer, Reese & Thompson, in press).

Classifications of the Various Effect Sizes

Various researchers characterize the magnitude-of-effect estimates in several different ways: estimates of the magnitude of the effect, estimates of the magnitude of the experimental effect, estimates of explained variance, effect size estimates, estimates of the strength of relation, or estimates of the strength of association, effect size estimates will be the term used here (Snyder & Lawson, 1993). However, it is important to realize that these terms are used interchangeably within the literature. The phrase "effect size" can be used to mean "the degree to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false" (Cohen, 1988). Effect size includes mean difference indices, estimated effect parameter indices, and standardized differences between means; therefore, this category consists of those measures that involve directly examining differences between means (Snyder & Lawson, 1993).

Effect size is a name given to a large number of indices that measure the magnitude of a treatment effect. Effect size can be classified into two general families, standard differences and variance-accounted-for measures of strength of association. Within both families, several different choices of effect sizes are available (Snyder & Lawson, 1993).

Standardized Differences

Standardized differences are defined as the standardized difference between two groups. There are several types of measurements that compute standard differences, such as Cohen's d, Glass' Δ, and Hedges' g. The

computational definition of standard difference is the experimental group mean minus the control group mean, divided by some estimated population standard deviation. Kirk (1996) summarizes Cohen's d, Glass' Δ, and Hedges' g quite thoroughly.

Cohen's d formula is as follows (μ - estimated population, σ - estimated population standard deviation):

$$\text{Cohen's } d = \mu_1 - \mu_2/\sigma$$

Cohen's d is the most popular of the three effect size measures discussed. Cohen's d expresses the size of the population treatment effect in units of the common population standard deviation, and Cohen provided guidelines for interpreting the magnitude of d. A medium effect of .5 was possible to see with the naked eye, and seen as the average size of observed effects in various fields. A .2 and .8 are both equally distanced from .5 on opposite sides and are considered low, but not trivial, and high effect, respectively. This guideline of interpretation, or operational definition, turned d into a much more usable statistic. Cohen's d was much more useful in the fact it could estimate the sample size necessary to detect small, medium, and large effects and to assess the power of a research design to detect various size effects. Correlation coefficients, regression coefficients, differences between correlation coefficients, proportions, differences between proportions, contingency table data, and differences among means in analyses of variance are also interpreted from Cohen's d (Kirk, 1996).

Glass' Δ effect size formula is as follows (Y_E - Experimental Group Mean, Y_C - Control Group Mean, S_C - Sample Standard deviation of Control Group):

$$\text{Glass' } \Delta = Y_E - Y_C / S_C$$

Glass' Δ used the effect size concept while working on meta-analysis data. Glass used a similar formula as d; however, he replaced the pooled standard deviation across groups with the sample standard deviation of only the control group. Glass believed that if there were several experimental groups, pairwise pooling of those standard deviations would result in a different standard deviation for each experimental-control contrast. Different effect size values due to the standard deviations of the contrasts differences would be the direct result of size difference between experimental and control means being the same size (Kirk, 1996).

Hedge's g effect size formula is as follows (Y_E - Experimental Group Mean, Y_C - Control Group Mean, S_{Pooled} - Pooled standard deviations of experimental mean and control mean):

$$\text{Hedge's } g = Y_E - Y_C / S_{Pooled}$$

Kirk (1996) describes Hedges g as only slightly different from the other two approaches of effect size. Hedges pooled the standard deviations of the experimental groups with that for the control group to obtain one standard deviation for all contrasts. His pooled population estimator is the same as the within-groups mean square in analysis of variance

Variance-accounted-for Measures of Strength of Association

Alternatively, variance-accounted-for effect size can be computed in all studies due to all analyses being correlational (Thompson, 1991). Variance-accounted-for measures of strength of association is defined as the variance-accounted-for squared correlation between the independent and dependent variables. Measurements in this category may be interpreted directly, or corrected. Such measurements that compute strength of association that are uncorrected effect size measurements are R^2 and eta squared, while uncorrected effect size measurements are omega squared and epsilon squared (Rosnow & Rosenthal, 1996; Thompson, 1996). Thompson (1996) explained that corrected effect size measurements may be used to estimate and adjust for the positive bias associated with smaller sample sizes, using more variables, and/or smaller population effects.

Sample size has been shown to influence statistical significance, which shows statistical significance can be manipulated by changing the sample size by one participant. Therefore, result interpretations should include explicit analyses when statistically nonsignificant results can be turned into statistically significant results simply by changing the sample size (Thompson, 1988). Variance-accounted-for statistics are the types of explicit analyses used in this type of situation within research result interpretation.

The most simple of variance-accounted-for measures of strength of association, also known as positively biased magnitudes of association

estimates, are eta squared (ANOVA) and R^2 (regression). The formula for eta squared and R^2 is as follows (SS - Sum of Squares):

eta squared and R^2 $SS_{explained}/SS_{total}$

Eta squared and R^2 can be expressed as the ratio of explained variance to total variance. They are positively biased because they tend to overestimate systematically the proportion of variability that might be explained in the population or in future samples (Snyder & Lawson, 1993).

Snyder and Lawson (1993) reported Stephens (1992) explanation of reasons for overestimations and biased estimates. The overestimates actually result from the mathematical maximization principle ("least squares") operating in all general linear model analyses. When sample results are analyzed, the linear combination of Xs that is maximally correlated with some Y is sought, and minimizing the sum of squared errors is equivalent to maximizing the correlation between X and Y scores. Therefore, any sample-specific idiosyncratic variation in the study samples that arise from the sampling error will cause a positive bias, and even if there is not a systematic relationship between X and Y in the population, R^2 or eta squared is not likely to ever equal zero.

O'Grady (1982) showed how the bias within eta squared and R^2 may vary depending on factors such as reliability of scores on the measurement instruments, research questions posed, sample size, number of predictor or independent variables under investigation in a particular study, heterogeneity

of the study sample, and type of design used to investigate a particular research question.

Due to the many areas noted for influence of possible bias in these effect size estimates, "corrected" measurements, also known as unbiased effect size estimates, have been developed. Corrected effect size measurements, omega squared and epsilon squared, differ from eta squared and R^2 in that they adjust for the sampling error present in both a given present study and future studies. The formula for omega squared is as follows (SS - Sum of Squares, v - number of levels in a factor, MS_{error} - ,):

$$\text{Omega squared} \quad SS_{explained} - [(v-1) * MS_{error}]/SS_{total}+MS_{error}$$

The formula for omega squared is as follows:

$$\text{Epsilon squared} \quad SS_{explained} - [(v-1) * MS_{error}]/SS_{total}$$

This adjustment in sampling error results in the "shrinkage" of the original estimates for future samples. The types of generalizations the researcher wishes to make plays a major role in whether the bias correction formulas designed to estimate measure of association strength is to be used in the result interpretations (Snyder & Lawson, 1993

Snyder and Lawson (1993) described in more depth the various formulas to choose from when estimating the association strength, or effect size. The different designs discussed by Snyder and Lawson are fixed- versus random-effect design models, univariate versus multivariate magnitude-of-effect estimates (for multivariate cases only), and equivalent estimates from varying perspectives of the general linear model.

Once researchers recognize the usefulness of effect size estimates, results will involve more informed analyses of data, and a more applicable flavor in the real world. Practical significance has made it possible to take result interpretations and apply them to the real world, whether or not statistical significance was found. Null hypothesis significance testing has been seen as essential in the world of research for the past 70 years, but has finally been recognized as somewhat limited in result interpretation. Due to the controversy of resistance to accepting the limitations surrounding null hypothesis significance testing supplemental procedures have been developed. Now that there is a better understanding of the importance of reporting some type of effect size, APA has been working on changing some of the guidelines, and thirteen journals now require effect size reports, while some journals strongly recommend effect size reports (Kirk, 1996; Snyder & Lawson, 1993).

New and Upcoming Guidelines for Effect Size in APA

Following decades of criticisms of statistical significance testing practices (cf. Carver, 1978; Cohen, 1994; Meehl, 1978; Schmidt, 1996; Thompson, 1996), APA now "encourages" effect size reporting in journal articles. Carver (1978) defined statistical testing as something more like fantasy than fact, and argued that statistical significance testing should be given little space in the results section. Carver said that if statistical significance testing was eliminated, a way of collecting and analyzing data that provides convincing evidence needed to replace it. Cohen (1994) said that statistical testing does not tell us what we want to know, but we want to know so badly what we are

looking for, that we accept it nevertheless. The alternative to the constant controversy is to report both, if anything at all in the result interpretation. Several journals began to see the importance, not statistical significance, of reporting effect size. At least 13 journals now "require" such reports (e.g., Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994): Career Development Quarterly, Contemporary Educational Psychology, Educational and Psychological Measurement, Journal of Agricultural Education, Journal of Applied Psychology, Journal of consulting & Clinical Psychology, Journal of Early Intervention, Journal of Experimental Education, Journal of Learning Disabilities, Language Learning, Measurement and Evaluation in Counseling and Development, The Professional Educator, and Research in the Schools.

Another important area of interest to those wanting to publish research is the recently published report of the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999), which will be incorporated into the 2001 revision of the APA publication manual. Soon all social science journals will be requiring effect size reports. The Task Force emphasized, "Always provide some effect-size estimate when reporting a p value" (p. 599, emphasis added). Later the Task Force also wrote,

> Always present effect sizes for primary outcomes....It helps to add brief comments that place these effect sizes in a practical and theoretical context....We must stress again that reporting and interpreting effect sizes

> in the context of previously reported effects is essential to good research (p. 599, emphasis added).

In summary, there are a number of ways one can compute an effect size statistic as a part of data analysis. There is no concept of "one-size fits all" (Thompson, 1999), so it is up to the discretion of the informed researcher to choose the index best suited for a particular research endeavor. Cohen (1994) closes his famous article The Earth is Round (p<05) by placing full responsibility on the researcher by saying,

> ....we have a body of statistical techniques, that, used intelligently, can facilitate our efforts. (p. 1002)

However, choosing a supplemental statistic such as effect size, along with statistical testing has now become necessary that such a statistic always be included to enable other researchers to carry out meta-analyses and to inform judgment regarding the practical significance of results. This includes the ability to replicate research, which also falls under the law of Cohen's The Earth is Round (p < .05) in that "we must finally rely, as have the older sciences, on replication. (p. 1002)."

References

Anderson, D.R., Burnham, K.P., & Thompson, W. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management, 64, 912-923.

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.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Kieffer, K.M., Reese, R.J., & Thompson, B. (in press). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. Journal of Experimental Education.

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

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61(2).

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.

Murhpy, K.R. (1997)Editorial. Journal of Applied Psychology, 82, 3-5.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemporary Educational Psychology, 25, 241-286.

Rosnow, R.L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychological Methods, 1, 331-340.

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

Shibley Hyde, J. (2001). Reporting effect sizes: the roles of editors, textbook authors, and publication manuals. Educational and Psychological Measurement, 61(2).

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

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

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

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sized and the bootstrap. Invited address presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

Thompson, B. (in press). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development.

Thompson, B., & Kieffer, K.M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. Research in the Schools, 7(2), 3-10.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. Educational and Psychological Measurement, 61(2).

Wilkinson, L., & 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]

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

REPRODUCTION RELEASE

(Specific Document)

TM032357

I. DOCUMENT IDENTIFICATION:

Title: AN INTRODUCTORY SUMMARY OF VARIOUS EFFECT SIZE CHOICES	
Author(s): SUSAN CROMWELL	
Corporate Source:	Publication Date: 2/1/01

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

SUSAN CROMWELL

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: Susan Cromwell	Position: RES ASSOCIATE
Printed Name: SUSAN CROMWELL	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: 979/845-1335 Date: 1/25/01

OVER