Advances in Health Sciences Education **9:** 241–249, 2004.

Methodologist's Corner

A Visitor's Guide to Effect Sizes ⋆

Statistical Significance Versus Practical (Clinical) Importance of Research Findings

MOHAMMADREZA HOJAT[1]* and GANG XU[2]
[1]*Center for Research in Medical Education and Health Care and Department of Psychiatry and Human Behavior, Jefferson Medical College of Thomas Jefferson University;* [2]*The International Cultural Exchange Institute, Eastern Region of Anhui University of China (*Author for correspondence, 1025 Walnut Street, Jefferson Medical College, Philadelphia, PA 19107, USA; E-mail: Mohammadreza.Hojat@Jefferson.edu)*

Abstract. Effect Sizes (ES) are an increasingly important index used to quantify the degree of practical significance of study results. This paper gives an introduction to the computation and interpretation of effect sizes from the perspective of the consumer of the research literature. The key points made are:

1. *ES* is a useful indicator of the practical (clinical) importance of research results that can be operationally defined from being "negligible" to "moderate", to "important".
2. The *ES* has two advantages over statistical significance testing: (a) it is independent of the size of the sample; (b) it is a scale-free index. Therefore, *ES* can be uniformly interpreted in different studies regardless of the sample size and the original scales of the variables.
3. Calculations of the *ES* are illustrated by using examples of comparisons between two means, correlation coefficients, chi-square tests and two proportions, along with appropriate formulas.
4. Operational definitions for the *ES*s are given, along with numerical examples for the purpose of illustration.

Key words: clinical significance, effect size, medical education research, practical importance

I. Estimates of Effect Sizes

Effect size can be considered an index of the *extent* to which the research hypothesis is true, or the *degree* to which the findings have practical significance in the study population. In other words, effect size is an index that quantifies the degree

⋆ Based on our experiences in teaching statistics and research methodology to medical and other health professions students, we believe that the first step in evolving to a practitioner of research is to become an informed consumer of research. A better-informed individual can accept or reject the research findings with a better critical view. In this article, we will describe, in a non-technical language, the procedures for calculating the effect size estimate and determining the practical (clinical) significance of research findings as opposed to the regularly reported statistical significance of findings. The conceptual and technical details in this article are not certainly sufficient for the practitioners of research, but may help the *consumers* of research to better understand and hopefully enable them to critically evaluate the research findings.

to which the study results should be considered negligible, or important, *regardless* of the size of the study sample.

For the purpose of operational definitions of the magnitudes of treatment differences (usually one or more experimental groups compared to a control group), or group differences in non-experimental designs, and for comparability of findings in different studies (e.g., meta analyses) it is desirable to have a "scale-free" effect size index. Estimates of effect sizes are calculated as standardized differences to serve that purpose. Therefore, effect size estimates can be used to compare treatment effects for different variables in the same study, or for the same or different variables across different studies, regardless of the study sample size and the original scales of the variables. These are important characteristics of any estimate of effect size that also have important implications in meta-analytic studies.

Because of these two important advantages of the effect size estimates (independent of sample size, and scale-free characteristic), some professional research journals recently began to recommend, and some require, that the authors report the effect size estimates of the findings in their submitted articles. In the latest publication manual of the American Psychological Association (APA, 2001), for instance, authors are encouraged to report effect size values in any empirical study.

Effect size estimates can be calculated for many different statistical indices. We have chosen the following three topics because of their frequent use in medical education research: mean differences (*t*-test), measures of association (correlation coefficient, chi-square), and the difference between two proportions.

1. Effect size estimate for mean differences

In this section, we will discuss effect size estimates for the differences between two means. We shall use ES to represent effect size estimate from this point forward. The typical inferential statistical method used to examine the statistical significance of the difference between two means is the *t*-test. Three different cases are discussed below.

a. Comparing Two Independent Samples

Comparisons of the means for two independent groups, such as an experimental and control groups, are commonly reported in the literature. Calculation of the *ES* in this case is simply the difference between the means of experimental (M_e) and control (M_c) groups divided by the standard deviation for the control group (σ_c).

$$\boxed{ES = |M_e - M_c|/\sigma_c}$$

This ES is sometimes referred to as the *"Glass's effect"*. We should point out that there is a dispute among experts about the appropriate denominator in the formula used in experimental designs (see Morris and DeShon, 2002).[1]

One point deserves attention. For the sake of simplicity, we report the absolute value of the *ES* throughout this article, without making any distinction between

directional/non-directional hypotheses (one-tail/two-tail tests). Researchers should take the algebraic sign of the *ES* into consideration when the direction of group differences is a consideration (one-tailed test). However, the direction of a hypothesis does not influence the magnitude of the *ES*.

The aforementioned formula is applicable in situations in which experimental and control groups are used. In situations employing quasi-experimental and ex post facto designs, the means for two groups are usually compared without necessarily assigning experimental or control groups status to the groups (e.g., men compared to women, generalists compared to specialists, etc.) In such cases, a minor modification to the previous formula should be made in the denominator. The pooled within group standard deviation (σ_{pooled}) should be used instead of the σ for the control group.

$$ES = |M_{\text{group1}} - M_{\text{group2}}|/\sigma_{\text{Pooled}}$$

Where, σ_{pooled} is the standard deviation that can be calculated directly from the combined data for both groups, or by using the following formulas: $\sigma_{\text{pool}} = \sqrt{}\,[(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2] / (n_1 + n_2 - 2)$; and in the case of equal sample size in the two groups: $\sigma_{\text{pooled}} = \sqrt{}\,(\sigma_1^2 + \sigma_2^2) / 2$; where σ_1^2 and σ_2^2 are variances for group 1 and group 2, respectively, and n_1 and n_2 are the sample sizes of the respective groups. This index is sometimes referred to as "Cohen's *d*".

Effect sizes can range from negative to positive infinity, but in practice, usually stay within +/–1. To help interpret ES values, Cohen (1987) has classified *ES values* into three different categories of "small", "medium", and "large", and operationally defines the typical magnitude of the *ES* in each of these categories for different statistics. In the case of comparing two means (Cohen, 1987, p. 40), Cohen's operational definitions are:

ES ≅ 0.20 (SMALL, negligible practical importance)
ES ≅ 0.50 (MEDIUM, moderate practical importance)
ES ≅ 0.80 (LARGE, crucial practical importance)

In clinical research, "clinical" instead of "practical" importance is usually used (see Numerical Example #1).

b. Comparing Matched Samples or Repeated Measures

In this case the means being compared are from: paired (matched) groups; or, the means are from two measurements taken from the same group (e.g., repeated measures such as a pre-test post-test design). Although for the sake of simplicity we recommend to use the baseline (pre-test) σ in the denominator of the formula, it should be mentioned that different approaches have been suggested in calculating the denominator in repeated measure designs (see Dunlap, Cortina, Vaslow and Burke, 1996).[2] *ES* is calculated by the same formula used for independent samples (see Numerical Example #2).

c. Comparing the Mean of a Sample with that of Its Respective Population

The case of comparing a sample mean with the population mean is not frequently reported in the literature. Generally in this case we want to know if the sample can be considered to come from the given population, although there are times when the desire is to test whether the population mean is equal to a hypothetical value. To compare the mean for a sample (M_s) with the population mean (M_{pop}). *ES* is calculated by the following formula:

$$ES = |M_s - M_{\text{pop}}|/\sigma_{\text{pop}}$$

Where σ_{pop} is the population standard deviation (see Numerical Example #3).

2. Effect size estimate for measures of association

In this section we discuss the calculation of the *ES* for the index of association between continuous measures (product moment, or Pearson correlation coefficient, *r*), and for the index of association between discrete measures (chi-square, χ^2).

a. Effect Size Estimate for Pearson Correlation Coefficient

The Pearson correlation coefficient, *r*, is one of the most frequently reported statistics in research. Fortunately, *r* is a scale-free statistic. It is a standardized index because r^2 is the proportion of common variance (or overlap) between the variables being correlated regardless of the size of the sample and the original scales of the correlated variables. For example, suppose that the magnitude of the correlation between two variables is 0.60 (e.g., scores of Step 1 of the United States Medical Licensing Examinations [USMLE], and scores on Biological Sciences Sections of the Medical College Admission Test [MCAT]). The proportion of shared variance (overlap) between the two variables is 36% ($r^2 = 0.60^2 = 0.36$). This conclusion will be true regardless of the sample size and the original scales of the two variables Therefore, the magnitude of the correlation coefficient is itself an effect size estimate. According to Cohen (1987, pp. 79–80), the operational definitions of the *ES* for correlation coefficients are as follows:

$ES = r \cong 0.10$ (SMALL, negligible practical importance)
$ES = r \cong 0.30$ (MEDIUM, moderate practical importance)
$ES = r \cong 0.50$ (LARGE, crucial practical importance)

b. Effect Size Estimate for Comparing Two Correlations

Calculation of the ES for the difference between two correlations is not as straightforward and simple as that for the mean difference. The reason is that the variance of the correlation is dependent on the value of the correlation. The implication of this dependence is that the *same* magnitude of difference between two correlations cannot always be considered equal on the correlation scale. For example, the

difference between $r = 0.75$, and $r = 0.95$ is 0.20, which is equal in magnitude to the difference between $r = 0.30$ and $r = 0.50$. But the statistical power to detect the difference between $r = 0.75$ and $r = 0.95$ is higher than that between $r = 0.30$ and $r = 0.50$, despite the fact that the difference between the two correlations that are being compared is the same. The solution to this issue is to transform the correlations to values on a new scale with equal interval characteristics. The transformation, called the Fisher Z transformation (Z), has the following form:

$$Z = 0.50 \log_e[(1 + r)/(1 - r)]$$

($\log_e$ is the natural logarithm or logarithm to base e. Tables are available for the transformation of correlation coefficients to their corresponding Z values. For example, an abridged version of such tables can be found in Cohen (1987, p. 112, Table 4.2.2); a more detailed table can be found in Owen (1962, pp. 511–512, Table 19.2).

For calculating the effect size estimate between r_1 and r_2, the following formula is used:

$$\boxed{ES = Z_1 - Z_2}$$

Operational definitions of the *ES* for differences between two correlations, as suggested by Cohen (1987, pp. 115–116) are as follow:

ES $\cong$ 0.10 (SMALL, negligible practical importance)
ES $\cong$ 0.30 (MEDIUM, moderate practical importance)
ES $\cong$ 0.50 (LARGE, crucial practical importance)

(see Numerical Example #4).

c. Effect Size Estimate for Chi-Square

The statistical significance of the association between discrete variables can be tested by the chi-square (χ^2) test. The *coefficient of contingency*, or C, is a widely used measure of association between discrete measures in contingency tables that can be derived from χ^2. *C is*, calculated by the following formula:

$$C = \sqrt{\chi^2/(\chi^2 + N)}$$

Where N is the total number of observations in the contingency table from which the χ^2 is calculated. The *ES*, can be calculated by using the contingency coefficient or directly from the χ^2 value (for 2×2 contingency table) in the following formulas:

$$\boxed{\begin{array}{l} ES = \sqrt{C^2/(1 - C^2)} \\ ES = \sqrt{\chi^2/N} \text{ for } 2\times2 \text{ contingency tables} \end{array}}$$

Operational definitions for interpreting the *ES* of χ^2, as suggested by Cohen (1987, pp. 224–225) are as follow:

$ES \cong 0.10$ (SMALL, negligible practical importance)
$ES \cong 0.30$ (MEDIUM, moderate practical importance)
$ES \cong 0.50$ (LARGE, crucial practical importance)
(see Numerical Example #5).

3. Effect size estimate for proportions

In determining the *ES* for comparing two proportions we must transform the proportion, *P*, to a new value, called *phi* (ϕ) because, like correlation coefficients, the variance of a proportion [P(1 – P)] is dependent upon the value of the proportion. The implication of this dependence is that while the difference between proportions 0.10 and 0.25 (0.15) is the same as that between 0.80 and 0.95, the statistical power to detect these differences between the first pair and the second pair is different in spite of identical differences in proportions. The nonlinear transformation of *P* to ϕ adjusts for the non-equal unit of detectability of the proportions. The following formula is used to transform *P* to the corresponding ϕ:

$$\phi = 2(arcsin\sqrt{P})$$

Here, $arcsin\sqrt{P}$ is the inverse trigonometric function of $\sin\sqrt{p}[\sin^{-1}\sqrt{P}]$. Tables are available for transforming *P* to ϕ values. (See Cohen, 1987, p. 183, Table 6.2.2 for an abridged version, and for a more detailed table in Owen (1962, pp. 296–303, Table 9.9).

The *ES* for the difference between two proportions (P_1 and P_2), can be determined by calculating the difference between their corresponding ϕs:

$$\boxed{ES = |\phi_1 - \phi_2|}$$

Operational definitions of the magnitude of *ES* for differences between two proportions, according to Cohen (1987, pp. 184–185) are as follow:

$ES \cong 0.20$ (SMALL, negligible practical importance)
$ES \cong 0.50$ (MEDIUM, moderate practical importance)
$ES \cong 0.80$ (LARGE, crucial practical importance)
(see Numerical Example #6).

Notes

[1] For example, Hedges (1982) suggests that instead of using the standard deviation of the control group, pooled within-groups standard deviation should be used in the denominator.

[2] For example, in calculating effect size in pretest-posttest designs, Cohen (1987) suggests that the standard deviation of pre-posttest score differences be used in the denominator. Morris (2000) recommend using the standard deviation of the pretest (baseline) scores for that purpose. In pre-posttest designs involving an experimental and a control group, Gibbons, Hedeker and Davis (1993) suggest to use the standard deviation of the pretest-posttest score differences of the experimental group in the formula; Hedges (1981) recommend that the effect size in these designs can be calculated by posttest means difference between the experimental and control groups, divided by the pooled

posttest standard deviation (also see Morris and DeShon, 2002). For estimating the clinical significance changes resulting from psychotherapy, Jacobson and Truax (1991) recommend calculating the mean of pre-therapy-post-therapy score differences divided by the standard error of the differences between the two test scores. They call this effect size as an index of reliable change (RC).

Numerical Examples

Numerical Example #1: A medical school faculty wanted to test if students in an "active" learning program (Group 1: students were encouraged to find answers to the issues by independent study) could perform differently than their classmates in another program (Group 2: answers to the issues were given directly to students by their expert instructors). The students were randomly assigned into the two groups at the beginning of the second year of medical school. Scores of the two groups on a standardized measure of knowledge in basic medical sciences (Step 1 of the United States Medical Licensing Examination, USMLE) taken at the end of the second year were compared. The following results were obtained:

$$M_{\text{Group1}} = 211, \quad \sigma_{\text{Group1}} = 16.9$$
$$M_{\text{Group2}} = 196, \quad \sigma_{\text{Group2}} = 14.3$$

The difference between the two groups is statistically significant ($t = 2.14$, $p < 0.05$). Is this difference practically important?

$$\begin{aligned} ES &= |211 - 196|/\sqrt{(16.9^2 + 14.3^2)/2} \\ &= 15/\sqrt{(285.6 + 204.5)/2} \\ &= 15/\sqrt{409.1/2} \\ &= 15/\sqrt{245.1} \\ &= 15/15.7 \\ &= \mathbf{0.96}. \end{aligned}$$

An effect size of this magnitude (0.96), according to the aforementioned operational definitions, should be considered to be practically important to a large degree. Therefore, active learning could increase students' performance on a standardized examination to a degree that can be considered of crucial practical importance.

Numerical Example #2: A group of medical students were shown a short video on how a patient and her family react to terminal illness. Students were given an empathy test before and after viewing the video. The following statistics were calculated:

$$\begin{aligned} M_{\text{pre-test}} &= 19.8 \\ M_{\text{post-test}} &= 20.9 \\ \sigma_{\text{p}} &= 1.9 \text{ (standard deviation of the pre-test scores)} \end{aligned}$$

The difference is statistically significant ($t = 2.4$, $p < 0.05$).

$$ES = |19.8 - 20.9|/1.9 = |\mathbf{0.58}|$$

With and *ES* of this magnitude we may conclude that viewing the video increased the empathy scores of the students to a degree considered of medium practical importance.

Numerical Example #3: The Dean of a medical school wanted to know if there was a practically important difference between the average score of 197 on Step 2 of the USMLE obtained by students in his medical school compared to the national average on this examination (assuming that the national average = 200, and standard deviation of the examination at national level is 20). The difference is statistically significant ($p < 0.05$)

$$ES = |197 - 200|/20 = |\mathbf{0.15}|$$

This is a small *ES*, and therefore the difference on Step 2 scores between students in this particular medical school and all candidates nationally Although statistically significant, but is not of practical importance.

Numerical Example #4: A psychologist wants to test the hypothesis that "birds of the same feather fly together for a longer time". She randomly selects a group of happily married husbands and wives, and another group of divorced couples matched for relevant variables (e.g., age, education, number of children, etc.). Couples in both groups complete a scale of attitude toward premarital sex, marriage and the family. The correlations calculated for scores of husbands and wives are 0.65 for the happily married couples and 0.20 for the divorced couples. The difference between the two correlations is satistically significant ($p < 0.05$). Is the difference of practical importance?

$$\begin{aligned} \text{Z for group 1} &= 0.50 \log_e[(1 + 0.65)/(1 - 0.65)] = 0.50 \log_e(4.71) = 0.50(1.55) = 0.78 \\ \text{Z for group 2} &= 0.50 \log_e[(1 + 0.20)/(1 - 0.20)] = 0.50 \log_e(1.50) = 0.50(0.40) = 0.20 \end{aligned}$$

Therefore, $ES = 0.78 - 0.20 = \mathbf{0.58}$

An *ES* of this magnitude (> 0.50), as operationally defined before, is of Crucial practical importance.

Numerical Example #5: The association between board certification status (certified, not certified) and practice specialty (generalists, medical subspecialists, surgical subspecialists, hospital-based specialties) was statistically significant by chi-square test ($\chi^2 = 8.1$, $p < 0.05$; total $N = 800$). Is this association practically important?

$$\begin{aligned} C &= \sqrt{8.1/(8.1 + 800)} = \sqrt{8.1/(8.1 + 800)} = \sqrt{0.01} = 0.10 \\ ES &= \sqrt{0.10^2/(1 - 0.10^2)} = \sqrt{0.01/(1 - 0.01)} = \sqrt{0.01/0.99} = \sqrt{0.01} = \mathbf{0.10}, \end{aligned}$$

or $ES = \sqrt{8.1/800} = \mathbf{0.10}$

Based on the aforementioned operational definitions, while the association is statistically significant, it barely meets the criterion of being of practical importance.

Numerical Example #6: Dean of a medical school proudly reported to the Dean of another medical school in town that his school attracted 45 percent of female applicants from the state, significantly higher than 37 percent attracted to the other medical school across the rive. The Dean of the second medical school asked his research staff to find out if such a claim had any practical merit. The *ES* calculated for the two proportions by the research staff:

$$\begin{aligned} \phi_{45\%} &= 2(arcsin\ \sqrt{0.45}) = 2(arcsin\ 0.67) = 2(0.74) = 1.48 \\ \phi_{37\%} &= 2(arcsin\ \sqrt{0.37}) = 2(arcsin\ 0.61) = 2(0.65) = 1.30 \\ ES &= \phi_{45\%} - \phi_{37\%} = 14.8 - 1.30 = \mathbf{0.18} \end{aligned}$$

The research staff reorpted to their Dean that the effect size estimated of this magnitude was negligible. The Dean of the second medical school sent a short not with the statistical evidence to his counterpart, and suggested that he "cools off" a bit on his claim!

References

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th edition). Washington, DC: American Psychological Association.

Cohen, J. (1987). *Statistical Power Analysis for Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Owen, S.B. (1962). *Handbook of Statistical Tables*. Reading, MA: Addison-Wesley.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B. & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measure designs. *Psychological Methods* **1**: 170–177.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin* **70**: 245–253.

Gibbons, R.D., Hedeker, D.R. & Davis, J.M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics* **18**: 271–279.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related measures. *Journal of Educational Statistics* **6**: 107–128.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin* **92**: 490–499.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Jacobson, N.S. & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* **59**: 12–19.

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

Morris, S.B. (2002). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology* **53**: 17–29.

Morris, S.B. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods* **7**: 105–125.

Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw Hill.

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.