Effect Size vs Inferential Statistics

By Marilyn K. Simon and Jim Goes

Includes excerpts from Simon (2011), *Dissertation and Scholarly Research: Recipes for Success*. Seattle, WA: Dissertation Success LLC

Effect size is a descriptive statistic referring to the measurement of the strength of a relationship between variables under a specific situation (Wilkinson, 1999). For instance, if we have data on the salaries of male and female engineers working for a particular company, and we notice that on average male engineers make more money than females in this company, the difference between the salaries of men and women is known as the effect size. The greater the effect size, the greater the salary difference between men and women. A question remains about whether the effect size is statistically significant or not.

When evaluating an intervention, program, or treatment, a question often arises: *How much effect did this intervention/program/treatment have*? Program evaluations often access data from an entire population of interest, e.g., all participants in a professional development program for teachers. In such situations, data on the entire population are available and there is no need to use inferential testing because there is no need to generalize beyond the participants. In these situations, descriptive statistics and effect sizes may be all that is needed to determine the efficacy of the intervention/program or treatment.

An effect size does not make any statement about whether the apparent relationship in the data reflects a *true* relationship in a population. In that way, effect size complements inferential statistics such as p-values. Among other uses, effect size measures play an important role in meta-analysis studies that summarize findings from a specific area of research, and can be used in lieu of statistical power analysis (a technique in the design of experiments that helps

to determine how big a sample size *should* be selected for that experiment so that the results can be generalized to a larger population).

The concept of effect size is somewhat ubiquitous in many claims made by various companies regarding products or services. For example, a weight loss program may boast that Plan D leads to an average weight loss of 25 pounds in a month. Another example would be a certain additive to gasoline increases that is touted to increase fuel efficiency by 12 mpg. These are examples of *absolute effect sizes*, meaning that they convey the average difference between two groups (those who participate in a program or treatment and those who do not) without any discussion of the variability within the groups. For example, in Plan D an average loss of 25 pounds could indicate that every participant lost exactly 25 pounds, or half the participants lost 50 pounds and the other half *nada*.

The reporting of effect sizes facilitates the interpretation of the substantive, as opposed to the statistical, significance of a research result. Effect sizes are particularly prominent in social and medical research. Relative and absolute measures of effect size convey different information, and can be used complementarily. It is good practice to present effect sizes for primary outcomes, that is, outcomes that are expected to be analyzed relevant to the effects of an intervention/program/treatment under review.

According to Valentine and Cooper (2003), effect size can help determine whether a difference is *real* or more likely due to a change of factors. In meta-analysis, effect size is concerned with different studies that are combined into a single analysis. The effect size is often measured in three ways: Standardized mean differences; Odds Ratio; and Correlation Coefficient.

A *Standardized mean difference* is a summary statistic in a meta-analysis when more than one study was conducted to assess the same outcome but measure the outcome in a variety of ways. For example, if several studies were conducted to measure mathematics anxiety in high school students but different psychometric scales were used, it would be necessary to standardize the results of the studies to a uniform scale before they can be combined into

one summary analysis. In meta-analysis, *standardized* effect sizes are used as a common measure that can be calculated for different studies and then combined into an overall summary. ZCalc is an excel spreadsheet add on that is used to convert a standardized mean effect size (ES) into a z-score. Cohen's d and Hedge's are common measures of ES. A calculator to help you determine the effect size for a multiple regression study (i.e., Cohen's f^2), given a value of R^2 is found at: http://danielsoper.com/statcalc3/calc.aspx?id=5

An odds ratio is a relative measure of risk, indicating how much more likely it is that someone exposed to a certain factor or treatment under study will develop an outcome as compared to someone who is not exposed. An odds ratio of 1 indicates no association between exposure and outcome. Odds ratios measure both the direction and strength of an association. A free calculator to measure odds ratio can be downloaded at: http://www.all-freeware.com/results/odds/ratio/calculator

A *correlation coefficient* measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson. The correlation coefficient indicates the degree of a linear relationship between two variables. The correlation coefficient always lies between -1 and +1. A correlation of -1 indicates perfect linear negative relationship between two variables, +1 indicates perfect positive linear relationship, and 0 indicates lack of any linear relationship. According to Cohen (1988), when measuring the effect size using correlation coefficients, a correlation of $r \geq .5$ can be characterized as a large correlation, .3 = medium, and .1 = small. A free calculator to determine a correlation coefficient can be found at: http://www.alcula.com/calculators/statistics/correlation-coefficient/

When to use Effect Size

The following situations would benefit from reporting the effect size:

1. Program evaluation studies with less than 50 participants tend to lack sufficient statistical power (this is a determination of sample size to obtain a given level of significance) for detecting small, medium or possibly even large effects. In such situations, the results of significance tests can be misleading because they are subject to Type II errors (incorrectly failing to reject the null hypothesis). In these situations, it can be more informative and beneficial to use the effect sizes, possibly complimented with confidence intervals.
2. For studies involving large sample sizes (e.g., n > 400), a different problem occurs with significance testing because even small effects are likely to become statistically significant, although these effects may in fact be trivial. In these situations, more attention should be paid to effect sizes than to statistical significance testing.
3. When there is no interest in generalizing the results (e.g., we are only interested in the results for the sample). In these situations, effect sizes are sufficient and suitable to determine efficacy.

When evaluating an intervention/program/treatment the use of effect sizes can be combined with other data, such as cost, to provide a measure of cost-effectiveness. In other words, as noted by McCartney and Dearing (2002), how much bang (effect size) for the buck (cost) is an intervention/program or treatment worth?

Advantages and disadvantages of Using Effect Sizes

Some advantages of effect size reporting are that:

1. It tends to be easier for practitioners to intuitively relate to effect sizes (once the idea of effect size is explained) than to significance testing.
2. Effect sizes facilitate comparisons with internal and external benchmarks.
3. Confidence intervals can be placed around effect sizes (providing an equivalent to significance testing is used)

However, disadvantages of using effect sizes can include:

1. Most software packages tend to offer limited functionality for creating effect sizes.
2. Most research methods and statistics courses tend to teach primarily, or exclusively, classical test theory and inferential statistical methods, and underemphasize effect sizes. In response, there has been a campaign since the 1980s (see Wilkinson, 1999) to educate social scientists about the misuse of significance testing and the need for more common reporting of effect sizes.

.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). New York: Academic Press.

McCartney, K. & Dearing, E. (2002). Evaluating effect sizes in the policy arena. *The Evaluation Exchange*, 8(1).

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Wilkinson, L. (1999); APA Task Force on Statistical Inference. "Statistical methods in psychology journals: Guidelines and explanations". *American Psychologist* 54: 594-604. doi:10.1037/0003-066X.54.8.594.