Paper 1: What are effect sizes and why we need them?

Dr. Larry Hedges, Board of Trustees Professor of Statistics and Policy Research at Northwestern University

Effect sizes are a type of quantitative representation of the magnitude of relations, differences, or comparisons that are in some way meaningful in the research design to which they are applied. Two classic examples of effect sizes include the difference in group means and a regression coefficient relating Y with X. However, statistical significance, which has certain universal appeal because many analyses compute the p-value, is not a replacement for effect size. In addition, there are faults in using statistical significance as an index of differences. For example, large studies frequently find significant effects, small studies frequently fail to find significant effects, and it is not obvious that all statistically reliable effects are substantively important.

The reason that statistical significance is not an index of effect size is because the distribution of the test statistic depends on two things: (a) an *effect size* component (defined by substantively relevant population parameters and (b) a *design* component (including sample size). Effect sizes are a way of describing substantively important relations among population parameters in a way that is independent of the research design. This makes effect sizes from different studies comparable to one another.

This point can be demonstrated by examining the t-test, written as Test Statistic = (Effect Size) x (Design Component).

$$t = \left(\frac{\bar{Y}_1 - \bar{Y}_2}{S_{Pooled}} \right) \sqrt{\frac{n_1 n_2}{(n_1 + n_2) DEF}}$$

More importantly, the sampling distribution of the t-statistic is determined by the noncentrality parameter which has two components as well—the population parameter and the study design parameter.

$$\lambda = \left(\frac{\mu_1 - \mu_2}{\sigma} \right) \sqrt{\frac{n_1 n_2}{(n_1 + n_2) DEF}}$$

This equation highlights that the power and expected p-value are determined by both design and population characteristics.

The term "effect size," as used by Jacob Cohen, describes the part of the noncentrality parameters that are independent of sample size. For normal theory tests, these effect sizes tend to be standardized (since no test on means has power that is independent of the variance), which leads to effect size parameters like:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

A common question of standardized effect sizes is, why standardize quantitative representations of a standardized mean? First, standardization is mathematically "natural." More importantly, for

many cases of social measurements, there is no natural set of units. Scaling, such as test scores, is often arbitrary. Standardization represents one arbitrary, but potentially universal, choice of scale. Because scaling, in this sense, makes estimates of effect comparable across different study designs that use different kinds of outcomes, standardization promotes interpretability since it does not depend on any particular scale. Finally, standardized effects also have an interpretation in terms of overlap between distributions (i.e., Mahalanobis distance) and are interpretable even when comparing effects measured on "different" variables.

Standardized effect sizes can: express results independent of data collection design (as much as possible), express results independent of location and scale parameters of measurement scale, and ideally, express results in a way that is substantively interpretable. However, there are complications in choosing standardized effect sizes. One of them is that in multilevel populations, there is an ambiguity about how to standardize (i.e., what σ?). For example, if you have a population, as in education, where there are students nested in classrooms nested in schools it is not so obvious what standard deviation is the right one to use. Is it the classroom, school, total or some other standardization? Standardization by the total (not within group) variance is often sensible because it is an obvious standard, but it introduces slight technical complications and there are cases where it is arguably not the right variance to use for interpretability purposes (e.g., when comparing differences between schools or when subpopulations are of primary interest). There are often other choices about the population. Measurement error can influence both the definition and estimation of effect size. Should you consider the relation between true scores or observed scores? There are arguments in some cases to do one or the other that can be persuasive.

Another complication that arises is that sometimes research is done in restricted populations, but the interest may be in unrestricted populations. This issue occurs in a variety of settings, one of which is when researchers select extremes on one variable for research design, but want to infer to the relation in the whole population.

Standardized metrics are not the only effect size nor are they always the safest effect size. If natural measurement scales *are* available, then standardization may actually *reduce* interpretability. For example, expressing effects on physical measurements in a standardized scale makes no sense. The question of whether to standardize depends heavily on the kinds of outcomes that are used and whether or not the metrics in which the outcomes are measured are well understood.

Three families of effect sizes are widely used (a) the *standardized mean difference* family (including the **d**-index); (b) the *standardized regression coefficient* family (including the correlation coefficient); and (c) the *odds ratio* family (including the risk difference and risk ratio). Two other families of effect sizes that are used, but not discussed: (a) the response ratio family (often used in experimental ecology but not in social science) and (b) multiple degree of freedom effect sizes based on variance ratios (e.g., variance accounted for measures). They are seemingly simple, but deceptively difficult to interpret precisely. Effect sizes in each of these families can be (under certain assumptions) translated into one another, but sampling theory is better if the "natural" effect size is used for a given design.

Standardized Mean Difference Family

In population structure there are means in standard deviations symbolized by Greek letters and the effect size is the mean difference standardized by the population's standard deviation.

	Population	Sample
Data		
Means	μ1 μ2	$\bar{Y}_1$ $\bar{Y}_2$
SD	σ	S

Effect Sizes

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \qquad d = \frac{\bar{Y}_1 - \bar{Y}_1}{S}$$

Odds-Ratio Family

The mean outcome is measured as the proportion of cases having one of the two outcomes (the target outcome). Study data are proportions in groups one and two having the target outcome.

Population		Sample	
π1	π2	p1	p2

There are several ways to make an Effect Size by comparing $\pi 1$ with $\pi 2$

	Population	Sample
Risk Difference	$\Delta = \pi 1 - \pi 2$	$RD = p1 - p2$
Risk Ratio	$\rho = \pi 1 / \pi 2$	$RR = p1 / p2$
Odds Ratio	$\omega = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$	$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}$

The Standardized Regression Coefficient Family

The most commonly used effect size in this family is the correlation coefficient: *r.* This is the standardized regression coefficient when there are no other covariates.

Sample	Population
r	ρ

Interpretation of Effect Sizes

It is crucial to recognize that the interpretation of effect sizes is a judgment process. No statistical theory can make these judgments, but judgments need to be made within some normative framework. In his *Statistical Power* book, Cohen provided some guidelines about small, medium

and large effect sizes, which have been followed more rigidly than intended. In the psychological context, effect sizes should be considered in light of other characteristics to help researchers understand whether the effects are large enough to take seriously.

There are a variety of different kinds of normative data to compare and inform the interpretation of effect sizes. One example is the gap between socially relevant groups, such as Black-White, High SES-Low SES and Male-Female. A second example is indices of growth such as one year's average achievement growth and the rate of growth in a relevant period. Another example is to develop a normative understanding from collections of effect sizes of intervention studies (e.g., Lipsey and Wilson). Finally, the natural variation of relevant units such as inter-quartile range, probable error (median deviation from the mean), and distance between any relevant quartiles provides a mechanism for interpreting effect sizes.

In interpreting effect sizes, the level of analysis matters in multilevel populations: a difference that is large compared to variation at one level may be small in terms of variation at another level. For example, a mean difference that is small compared to between-student variation may be large compared to between-school variation. Thus, universal criteria for large or small effect sizes have significant limitations.

When reporting effect sizes, it is extremely important to include some notion of uncertainties (e.g., standard error or confidence intervals). The most useful effect size for any dataset may depend on the application. Therefore, it may be useful to report effects in several metrics (e.g., an odds ratio or a risk ratio for a given prevalence).