Effect Size Substantive Interpretation Guidelines: Issues in the Interpretation of Effect Sizes

Jeff Valentine and Harris Cooper, Duke University

When authors communicate the findings of their studies, there is often a focus on whether or not some intervention had the intended effect, and less attention to how much of an effect the intervention had. That is, it may be possible to state that some reading intervention increased reading scores more than the usual reading instructional techniques, but it is often more difficult for readers to determine *how much* of a difference the intervention made. Readers need to know if the intervention's effects are large or small, meaningful or trivial. These guidelines describe the strengths and limitations of several approaches to the assessing the magnitude of an intervention's effect. While there are few clearly right and wrong approaches, we do believe that some methods are better than others.

In general, there are three different approaches to assessing the magnitude of an intervention's effect: (a) assessing the statistical significance of the effects, (b) assessing practical significance based on the raw mean differences of experimental groups, and (c) assessing the relative size of the effects based on standardized estimates of effect size.

Statistical significance

The statistical significance of an outcome measure is sometimes used as a measure of effect size. Outcomes receiving a statistically significant result are treated as being big, important effects, while outcomes that turn out not to be statistically significant are treated as being unimportant.

The problem with this approach is that effects of the same size can sometimes be highly significant and at other times not significant. Alternatively, effects that don't matter much can be highly statistically significant, while effects that matter a great deal can be statistically not significant. This problem comes about because the test of statistical significance actually confounds two independent pieces of information: the magnitude of the intervention's impact (the effect size) and the size of the sample. Thus, statistical significance tells us very little (if anything) about the *practical significance* or *relative impact* of the effect size, and should not be used as a stand-alone measure of how much the intervention "matters."

Mean Differences between Comparison Groups

Sometimes, how much an educational intervention "matters" is easily understood, because the measurement metric is familiar to most people. For example, there might be an interest in determining if an intervention designed to improve graduation rates among at-risk youth has an impact on income at age 21. To discuss the relative magnitude of the effects, researchers need simply to let their readers know the mean differences in earned income at age 21 between participants and non-participants. We can then decide for ourselves whether the effect might be considered trivial or impressive.

Grade level equivalents

Another example of a relatively easily understood metric is *grade level equivalents* (GLE). GLE scores are based on the average score obtained on a standardized test by a norming group of students, given a specified amount of schooling. For example, a GLE of 3.2 on a math test means that the student is performing at a level equal to the average student in the norming group who had been in the third grade for two months. We can compare the average GLE score for a group of students who received some intervention with the average GLE with a group of students who did not receive the intervention and get a good intuitive sense about whether the difference is small or large.

Conversion to the test metric

Sometimes study results can be converted back into their former metric. For example, assume an intervention is designed to improve overall achievement and achievement is measured by scores on the Iowa Test of Basic Skills (ITBS). The researcher might conduct and report the results of a statistical test. The researcher might then state what the mean difference between the two groups was ("The intervention group scored, on average, fourteen points higher on the ITBS than the comparison group. This difference was statistically significant."). In this case it is relatively easy for individuals familiar with the score distribution of the ITBS to make a determination about the importance of the intervention's effect.

Binomial Effect Size Display

The Binomial Effect Size Display (BESD) is another way of presenting the results of a study in a manner that makes it easy to interpret the magnitude of the effects. For example, suppose an intervention was designed to improve the graduation rate among at-risk students. One way to carry out this study would be to identify a group of at-risk students, and randomly assign half of them to be in the dropout prevention program. The other half would serve as a comparison group. One way to analyze data from this study would be to create a BESD. A successful intervention might have results that look like this:

	Graduated	Did Not Graduate
Received Dropout Prevention Program	66	34
Did not Receive Dropout Prevention Program	34	66

The Binomial Effect Size Display (BESD) can be a dramatic and clear way to represent study results. We believe that the best use for BESD occurs when the data are similar to the example. That is, when there are two groups of participants and the outcome is measured dichotomously. Rosenthal & Rubin (1982) have demonstrated how BESD can be generalized to more common analytic situations. However, we believe these generalizations create conceptual difficulties for some audiences and should probably be avoided.

While the approaches outlined above are appealing, there are some things to be concerned about. First, as we note above, it is likely that individuals will need to be familiar with the distributional properties (e.g., how much variation is there around the mean score) of the outcome to make a reasonable judgment about the importance of the effect sizes. That is, the reader will need to have an understanding of the relative importance of, say, 15 points on the ITBS. If this is not the case, then the appropriate audience for this method is somewhat limited.

In addition, the unstandardized nature of this approach makes it ill-suited for comparing across measures, thus limiting its potential application. If readers are told that an intervention was associated with a 15 point improvement in ITBS scores and a 5% greater chance of graduating from high school, it is difficult to tell whether the intervention affects ITBS scores and graduation rates differentially. Thus, mean difference effect sizes can not be used to compare across outcomes.

The unstandardized nature of effect sizes based only on the mean differences between groups poses another problem when synthesizing the results of studies: there must be a sufficient number of studies that measure the outcome in the same way for the average effect size to be meaningful.

The above limitations suggest that using mean differences between groups to express the magnitude of an intervention's effect is not a satisfactory solution on its own. Instead, we recommend using raw mean differences, when available, and augmenting that presentation using metrics that can be more generally applied (even if they are less generally understood).

Comparing Effect Sizes Relative to One Another

All of the above approaches to interpreting effect sizes involve comparing the effect of an intervention against some explicit or implicit index of practical importance. In each case -- be it dollars, GLEs, high school graduates, or points on a standardized test -- there was either a stated or implied yardstick suggesting that, for example, people think saving $500 per pupil, raising GLEs by six months, or graduating 10% more students represents a practically significant program outcome, one that might be labeled "important" or "large".

What do we do when no such standard exists, or when the outcome measures used in studies varies from one to the next? One option is to use standardized effect sizes. In general, there are three families of effect sizes: the standardized mean difference, the correlation coefficient, and the odds ratio. Standardized effect sizes have the important advantage of allowing for comparisons between effects that are not measured on the same scale (as one might do in a research synthesis). For example, when conducting a research synthesis, some studies might report outcomes using the standardized mean difference, others might report outcomes using a correlation coefficient, while still others might report outcomes using an odds ratio. All of these different metrics can be transformed to a common metric, allowing the synthesist to combine the estimates. The primary disadvantage of standardized effect sizes is that they are less easily understood.

Standardized mean difference

The standardized mean difference statistic, referred to as *d* (Cohen, 1988), is a scale-free measure of the separation between two group means. Calculating *d* for any comparison involves dividing the difference between the two group means by either their average (pooled) standard deviation or by the standard deviation of the control group. This calculation results in a measure of the difference between the two group means expressed in terms of their common standard deviation or that of the untreated population. Thus, a *d* of .25 indicates that one-quarter standard deviation separates the two means. Due to the types of studies that the WWC is likely to review, we believe *d* will be the most common type of standardized effect size used by the WWC.

Correlation coefficient

The correlation coefficient, *r*, is a scale-free measure that assesses the degree to which two variables are related. The correlation coefficient is advantageous because relatively many people are familiar with it. One special case of the correlation coefficient is most likely to appear in studies reviewed by the WWC: the point-biserial correlation, r_{pb}, is a measure of association between one dichotomous variable (e.g., those that received an intervention compared to those that did not) and one continuous variable (e.g., grade point average).

Odds ratio

The odds ratio is another way to express the impact of an intervention, and is useful with two dichotomous variables. Formally, an odds ratio is calculated by dividing the odds of an event occurring (e.g., dropping out, obtaining a scholarship) in the group receiving an intervention by the odds of the event occurring in the group not receiving the intervention. For example, consider the dropout prevention program described above. Recall that of the 100 students receiving the program, 66 graduated. Of the 100 students not receiving the program, only 34 graduated. The odds of graduating for people in the program are Odds($_{\mathrm{PROGRAM}}$) = (66/34) and the odds of graduating for people not in the program are Odds($_{\mathrm{NO\ PROGRAM}}$) = (34/66). Thus the odds ratio for this program would be OR = (66/34) / (34/66) = 3.73. That is, the odds ratio suggests that the odds of graduating for the individuals receiving the dropout prevention program were 3.73 times greater than the odds of graduating for the individuals not receiving the prevention program.

Attempts to characterize the strength of standardized measures of effect size

There are several ways to characterize the strength of the standardized measures of effect size. Below we discuss the strengths and limitations of three of them.

Proportion of variance explained

One well-known method of characterizing the strength of a relationship involves calculating the proportion of variance explained. This is what happens when a researcher writes "The correlation between self-concept and achievement is *r* = +.30. Thus, self-concept accounts for 9% of the variance on achievement." The researcher arrived at the 9% number by squaring the correlation coefficient (.30 in the example). Squaring the correlation coefficient puts the estimate of the relationship in the context of the total variance in the outcome measure.

There are problems with using this metric as a measure of the strength of an intervention's impact. Most importantly, the proportion of variance explained often seems low and this can lead even experienced researchers to labeling efficacious interventions as ineffective (Rosenthal, 1984). We presume that more general audiences, such as the policymakers and the public, will experience these same problems to an even greater degree.

In addition, proportion of variance explained can be applied -- in a misleading fashion -- to comparisons involving more than two groups. For example, one study might compare reading intervention A, reading intervention B, and a control reading intervention. A second study might compare reading intervention B, reading intervention C, and a control intervention. A proportion of variance explained estimate can be derived from both studies. However, comparing or combining these estimates would be misleading, because the statistics were not focused on a single comparison between a reading intervention and a control group, making it impossible to know which interventions were different from which.

Cohen's benchmarks

Cohen (1988) attempted to address the issue of interpreting effect size estimates relative to other effect sizes. He suggested some general definitions for small, medium, and large effect sizes in the social sciences. However, Cohen chose these quantities to reflect the typical effect sizes encountered in the behavioral sciences as a whole -- he warned against using his labels to interpret relationship magnitudes within particular social science disciplines or topic areas. His general labels, however, illustrate how to go about interpreting relative effects.

Cohen labeled an effect size small if $d = .20$ or $r = .10$. He wrote, "Many effects sought in personality, social, and clinical-psychological research are likely to be small . . . because of the attenuation in validity of the measures employed and the subtlety of the issue frequently involved" (p. 13). Large effects, according to Cohen, are frequently "at issue in such fields as sociology, economics, and experimental and physiological psychology, fields characterized by the study of potent variables or the presence of good experimental control or both" (p. 13). Cohen suggested large magnitudes of effect were $d = .80$ or $r = .50$. Medium-sized effects were placed between these two extremes, that is $d = .50$ or $r = .30$.

A caution against using Cohen's benchmarks as generic descriptors of the magnitude of effect size is implied above. Because some areas, like education, are likely to have smaller effect sizes than others, using Cohen's labels may be misleading.

Proportion of distribution overlap

Cohen (1988) proposed another method for characterizing effect sizes by expressing them in terms of distribution overlap, called U_3. This statistic describes the percentage of scores in the lower-meaned group that are exceeded by the average score in the higher-meaned group. As an example, assume that high school students who do homework outperform students who don't do homework, and that the effect size is $d = +.20$. For this effect size, U_3 is approximately equal to 57. In this example, it would mean that if an average student left a high school in which all students did homework and moved to a high school in which no students did homework, then that students would move from the 50th percentile to the 57th percentile in achievement.

Discussion and Recommendations

The above discussion can be used to demonstrate the relative nature of comparing effect sizes. Suppose a study on the effects of homework revealed an average r of +.30. How should the relation's magnitude be interpreted? First, we could square the number to get a percentage of variance explained, but is 9% small or large? Second, we could create a BESD, but the outcome variables in the homework studies likely vary (some might be unit tests, others class grades) and are continuous, not dichotomous in nature.

Clearly then, interpreting homework's effects require us to be quite explicit in answering the question "Compared to what?" Interpretation will depend on the other relations chosen as contrasting elements. According to Cohen, homework's effect is a medium-sized behavioral science effect. Thus, compared to other relations in the behavioral sciences in general, this would be an average effect size, not surprisingly large or small. However, compared to clinical-psychological effects, this effect size may best be described as large, if we accept Cohen's suggestion that these relations are predominantly smaller than $r = .30$.

Comparing a specific effect size to effect sizes found in other disciplines or a discipline in general may be interesting but in most instances it is not very informative. The most informative interpretation occurs when the effect size is compared to other effects involving the same or similar variables. At the time Cohen offered his guidelines, comparing an effect size in a specialized topic area against a criterion as broad as "all behavioral science" might have been the best contrasting element available. Estimates of average effects for disciplines, subdisciplines, topic areas, or even single variables or operations were difficult to find. Today, these calculations are plentiful.

In sum, we believe that expressing an intervention's effect should be done using raw mean difference scores, when feasible. However, because of the limitations of this approach, this presentation should be accompanied by a standardized effect size (*d* or *r*, as appropriate) or an odds ratio. Because Cohen's (1988) labels give only the broadest interpretive yardstick for effect size, they should be used with appropriate caution. Because of its limitations, proportion of variance explained should not be used as an effect size in most cases. U_3 should be used when feasible.

References

Cohen, J. *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Acknowledgment

This document was reviewed by the WWC's Technical Advisory Group and the WWC's Steering Committee. Special thanks to Jesse Berlin, Larry Hedges, and Will Shadish for their helpful suggestions.

For citation purposes, please refer to this document as:

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