How to Select, Calculate, and Interpret Effect Sizes

Joseph A. Durlak
Loyola University Chicago

The objective of this article is to offer guidelines regarding the selection, calculation, and interpretation of effect sizes (ESs). To accomplish this goal, ESs are first defined and their important contribution to research is emphasized. Then different types of ESs commonly used in group and correlational studies are discussed. Several useful resources are provided for distinguishing among different types of effects and what modifications might be required in their calculation depending on a study's purpose and methods. This article should assist producers and consumers of research in understanding the role, importance, and meaning of ESs in research reports.

Key words clinical significance; effect size; meta-analysis; statistics.

The *Journal of Pediatric Psychology* (JPP) now requires authors to include effect sizes (ESs) and confidence intervals (CIs) in their manuscripts (http://www.oxfordjournals.org/our_journals/jpepsy/for_authors/editorial%20policy.pdf).

This article focuses on ESs and has a dual purpose: (i) to offer guidelines to producers of research regarding how to select, calculate, and interpret ESs obtained in their studies, and (ii) to help all consumers of research develop a better understanding of the role, importance, and meaning of ESs.

This article should be helpful because many pediatric studies contain small sample sizes, and it is important to know how this situation affects the calculation of ESs. Moreover, advances are continually being made in our understanding of the application and interpretability of ESs, and little guidance on these matters is available in most statistical texts, which do not devote much attention to ESs (Capraro & Capraro, 2002). As a result, many researchers are not well versed in incorporating ESs into their own work and most research reports in the social sciences do not contain ESs (Volker, 2006). This article focuses on several common types of ESs that cover many situations. Additional references are provided for circumstances that go beyond those described here.

This article is organized as follows. The first section discusses the necessity of ESs. The second section describes ESs commonly used in group designs and correlational studies, and the third focuses on the interpretation of effects. Finally, some concluding comments are offered. The appendix contains several equations for calculating different types of effects.

ESs are Necessary

The Task Force on Statistical Inference of the American Psychological Association recommended that researchers ''should *always* provide some ES estimate when reporting a *p* value'' (italics added, Wilkinson and APA Task Force on Statistical Inference, 1999, p. 599) and further emphasized that ''... reporting and interpreting ESs in the context of previously reported effects is essential to good research'' (p. 599). The fifth edition of the APA (2001) *Publication Manual* also stressed the importance of ESs by stating ''For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of ES or strength of relationship in your Results section'' (p. 25).

What is an ES? There are many different types of ESs but those discussed here provide information about the magnitude and direction of the difference between two groups or the relationship between two variables. An ES can be a difference between means, a percentage, or a correlation (Vacha-Hasse & Thompson, 2004). Researchers usually want to show there is a difference between the groups they are studying, or that some

All correspondence concerning this article should be addressed to Joseph A. Durlak, Department of Psychology, Loyola University Chicago, 6525 N. Sheridan Road, Chicago, IL 60626, USA. E-mail: jdurlak@luc.edu.

Journal of Pediatric Psychology 34(9) pp. 917–928, 2009
doi:10.1093/jpepsy/jsp004
Advance Access publication February 16, 2009

variables they are investigating are correlated. ESs provide this information by assessing *how much difference* there is between groups or *how strong* the relationship is between variables. In other words, ESs assess the magnitude or strength of the findings that occur in research studies. This is critical information that cannot be obtained solely by focusing on a particular *p*-value such as .05 (Thompson, 2006; Volker, 2006). *There is no straightforward relationship between a p-value and the magnitude of effect.* A small *p*-value can relate to a low, medium, or high effect. Moreover, as discussed later, *there is no straightforward relationship between the magnitude of an effect and its practical or clinical value*. Depending on the circumstances, an effect of lower magnitude on one outcome can be more important than an effect of higher magnitude on another outcome.

Relationship Between p-Values and the Magnitude of Effect

A *p*-value that is obtained in a research study is a function of both sample size and ES, and Thompson (2007) offered an excellent demonstration why effects should be calculated irrespective of their *p*-value. He presented the results of 10 studies with different sample sizes; only one of which reached a *p*-value of .05. This one study had an ES of −0.40 (an indication that the control group was superior to the experimental group; see below). Thompson noted that if attention was focused only on the study that achieved statistical significance, then the only interpretation possible would center around its negative finding. However, the average ES calculated across all 10 studies was +0.28, which was statistically different from zero and would result in a completely different interpretation.

Researchers Should Report ESs for All Outcomes Regardless of Their p-values

Pediatric study samples are often small. Snyder and Lawson (1993) have shown that even with a magnitude of effect as large as a *d* of .66, the addition of a single subject to a study with a small sample size can shift a *p* level above .05 to one below .05 without any change in the ES. A different scenario is possible with large samples sizes, where a small *p*-value might not yield a large effect. This is because with other factors held constant, increasing the sample size increases the likelihood of finding a statistically significant difference. Of course, researchers prefer results based on large samples, which increase confidence in the findings, but the point here is that we cannot predict the magnitude of an effect based only on the results of statistical significance testing.

In sum, including and interpreting ESs should not be viewed as just another editorial hurdle to pass for publication purposes, but as an essential component of good research. ESs achieve an important purpose because they help to assess the overall contribution of a study.

Once it is understood what an ES is in general terms, and what valuable information it provides, many ESs can be presented directly or easily calculated. What does become complicated are the nearly infinite number of ways that researchers can design their study and analyze their data which, in turn, may require some modification in the usual way of determining ESs. The following discussion focuses on the most common ES indices and research situations, but many additional user-friendly references are cited so that readers can gain further information that applies to their particular research scenarios. The description of different types of ESs begins first with those used in group designs.

ESs for Group Designs

This section covers three major types of effect that can be used in group designs: raw mean differences, standardized mean differences, and odds ratios (ORs).

Raw Group Differences

Readers may be unaware that a direct comparison of group means can serve as a useful ES. If Group A lost 10 lbs while Group B lost 5 lbs, then the ES is 5 lbs. If 20% of the intervention students graduated from high school, and only 5% of the controls did so, the ES is 15%. If the internalizing scores for two groups on the child behavior checklist differ by 10 points, this is the effect. In fact, the APA (2001) *Publication Manual* recommends the use of ESs on the original measurement scale whenever they are feasible.

Such straightforward group differences would serve as a very easy way to judge the magnitude of effect and to compare results across studies *if* studies all used the same metric for their outcomes. However, measurement strategies are often scaled differently so that comparing raw group differences across studies is not meaningful. The usual lack of direct comparability across measures in a research area thus requires an index that can be ''standardized,'' that is, expressed in a metric that is common across studies, as opposed to raw group differences that do not require standardization. This leads to a discussion of standardized mean differences (SMDs) as an index of effect.

Standardized Mean Difference

SMDs are usually used as ESs in group designs and in these situations, the ES is calculated by using the

difference between the post-test means in the numerator of the equation and using standard deviation units in the denominator, hence the term ''standardized'' mean difference. This standardization permits direct comparisons across studies using the same index of effect. A SMD of 0.50 based on outcome A from one study is directly comparable to a SMD of 0.50 calculated on that same outcome in another study.

The two most common SMD statistics are Hedges' *g* and Cohen's *d* [see Equations (1) and (2) in the appendix, respectively). There are some differences in how these statistics are calculated, but both are positively biased estimators of an ES when sample sizes are small. Therefore, it is important to correct for their upwards bias. The correction factor is in the second and third parts of Equations (1) and (2). Practically speaking, the correction amounts to a 4% reduction in effect when the total sample size is 20 and around 2% when $N = 50$ (Hedges & Olkin, 1985). Nevertheless, making this correction can be relevant for studies in pediatric psychology. Equations for converting Hedges' *g* into Cohen's *d*, and vice versa are included in the appendix.

SMDs are rounded off to two decimal places and the sign of the SMD typically represents a favorable result for intervention. A positive value would indicate the intervention group was superior to the control group on a positively oriented outcome measure (e.g., self-esteem or social skills) while a negative ES would represent superiority of the intervention on a negative-oriented outcome such as levels of depression. In the latter situation, a negative number would occur in the numerator of the first part of Equation (1) when a higher post mean of the control group assessing depression is subtracted from the lower post mean of the intervention group. In any case, a zero value for a SMD indicates no effect at all.

Theoretically, SMDs in control group designs can take any value, but 95% of all the mean ESs reported in 302 meta-analyses conducted in the social sciences have fallen between −0.08 and +1.08 (Lipsey & Wilson, 1993). This asymmetry around zero is not a statistical property of ESs, but reflects empirical findings that evaluated interventions are more likely to yield positive than negative ESs.

What to Do with More than Two Groups?

Many studies include more than one experimental group, for example, when an additional component is added to an intervention to test its benefit. A child social skills training condition constitutes one group and this condition is supplemented by a parent support condition which is a second experimental condition. There is also an untreated control group. A separate ES for each experimental condition can be calculated using Equations (1) or (2). An ES could also be calculated comparing the two experimental groups using either one in place of the control group in Equations (1) or (2). However, the latter ES (E group versus another E group) is likely to be lower in magnitude than the E versus C comparison because two interventions are being compared. Youth should benefit in each condition. Child therapy reviews have found that treatment to treatment comparisons yield ESs that can be up to one-third lower than those obtained in treatment versus control conditions (Kazdin, Bass, Ayers & Rodgers, 1990). Helpful advice for calculating SMDs for different research situations is provided by Lipsey and Wilson (2001).

What About Other Types of Intervention Designs?

Different designs require different calculations for ESs. For example, for one-group pre-post designs, the pre-group mean is usually subtracted from the post mean and divided by the SD at pre (Lipsey & Wilson, 2001). Researchers have used at least seven different types of effects for single subject designs (e.g., reversal or multiple baseline designs). The options are discussed by Busk and Serlin (1992) and Olive and Smith (2005). These authors favor an SMD for these designs that uses the intervention and baseline means and the baseline SD. The positive range of effects for *N* of 1 within-subject designs is usually much higher than for control group designs. Some meta-analyses of single subject research have shown that ESs can easily exceed 1.0 and can be as high as 11.0, depending on the method of computing ESs, the rigor of the design, and the outcome measure (Allison & Faith, 1995).

Finally, some authors employ cluster or nested designs, for example, students may be assigned to conditions based on their classroom or school or the intervention may occur in these settings. In these situations, the students are ''nested'' within classrooms/schools and the student data are not independent. While space does not permit detailed discussion of these designs, the What Works Clearinghouse's recommendations for calculating effects in nested designs and conducting appropriate significance testing should be followed (http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=19&tocId=1).

Some meta analytic reviews of control group designs using SMDs have calculated ''adjusted'' ESs by subtracting the pre-ES from the post-ES, with the pre-ES calculated in the same manner as the post-ES (e.g., Wilson, Gottfredson & Najaka, 2001; Wilson & Lipsey, 2007). Adjusted effects can be very important when group comparisons are

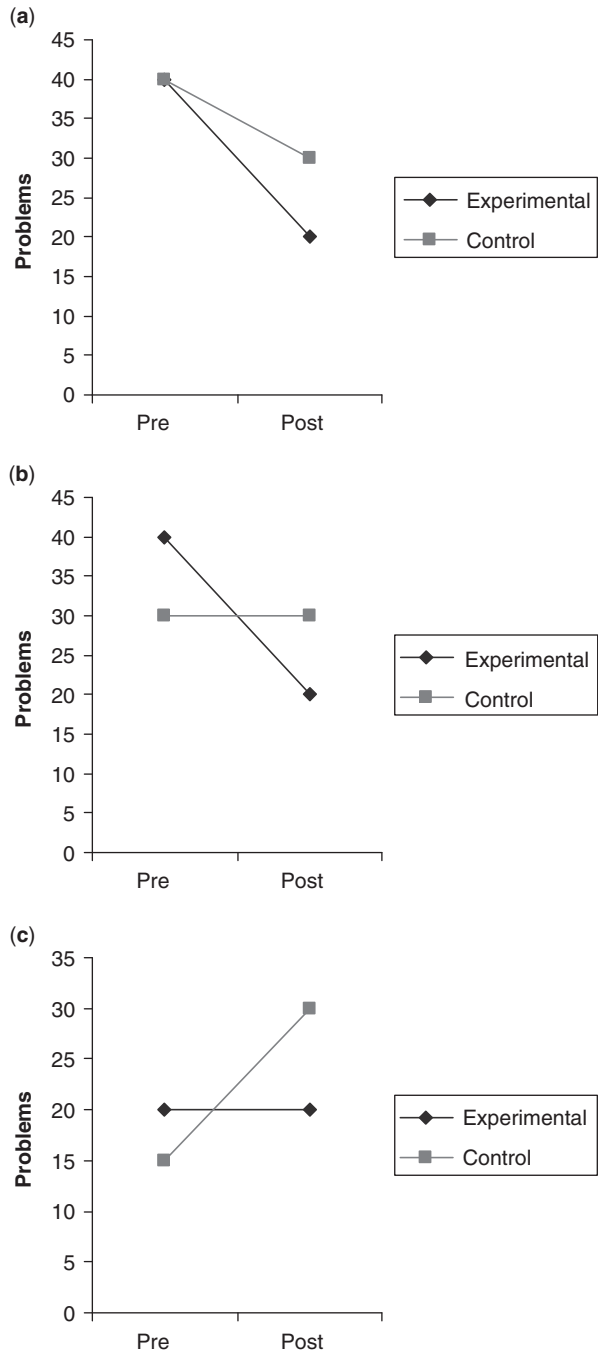

Figure 1. (A–C) Three depictions of the relative status of intervention and control groups over time

confounded by the lack of pre-test equivalence. For example, Figure 1 depicts three possibilities. In each case, the outcome is some measure of psychopathology so lower scores are preferred. In Figure 1a the two groups start out at exactly the same point; there is some improvement in both groups over time, but at post the intervention group compares favorably to the control group. In Figure 1b, the intervention group has more problems at pre than controls, but improves over time while the control group stays the same. In Figure 1c the intervention group once again has more problems than controls at pre but remains unchanged over time while the control group deteriorates from pre to post.

Many researchers would determine if the groups differed at pre by applying a *t*-test or one-way ANOVA, and if a *p* level of .05 was not reached, they might be tempted to conclude the two groups were initially equivalent. However, researchers should not mistakenly assume that the situation at pre in Figure 1a holds and there is no difference at all between groups (i. e., the pre-ES is zero). In fact, there will almost always be *some* between-group difference that might not be detected due to insufficient statistical power. There is likely to be more of a difference in quasi-experimental research than in randomized designs, but the expectation that randomized designs have established pre-test equivalence may not be realized, especially with smaller subject samples. For example, in a large scale meta-analysis of school-based interventions (Durlak, Weissberg, Taylor, Dymnicki & Schellinger, 2008) that included both quasi-experimental and randomized studies with a wide variation in samples sizes, less than 1% of the pre-ESs were truly zero. Admittedly, some pre differences might be negligible, but it can be worthwhile to assess their magnitude and then calculate and report adjusted ESs. Adjusted ESs should be reported along with the pre- and post-ES so as not to confuse readers.

Calculating an adjusted ES can have a major impact on the final result. Suppose the hypothetical pre- and post-ESs in Figure 1a are 0 and 0.20; in Figure 1b, they are −0.20 and +0.20 and in Figure 1c they are −0.20 and +0.20, respectively.

As a result, the adjusted ESs would be:

post-ES of 0.20 − a pre-ES of 0 = 0.20 for Figure 1a
post-ES of 0.20 − a pre-ES of (−0.20) = 0.40 for Figure 1b, and
post-ES of 0.20 − a pre-Es of (−0.20) = 0.40 for Figure 1c.

In other words, if the pre-ES was truly zero (a rare occasion), then the adjusted ES would be the same as the post-ES (0.20 for Figure 1a). But when it is not, ignoring the pre effect might make a huge difference Moreover, although the magnitude of the adjusted ES for Figures 1b and 1c are the same, they depict important differences in the relative adjustment of the two groups over time that are worthy of discussion in a report. For example, prevention sometimes is effective because it forestalls the

deterioration in adjustment that is observed in control groups over time as portrayed in Figure 1c. On other occasions, however, prevention can be effective because it results in a greater reduction of problems in the intervention groups than in controls (as in Figure 1b or in Figure 1a, but with some pre between-group differences). There are two different messages that can be drawn from such findings. In the former case, one might conclude that youth will worsen over time if we do not take effective preventive action; in the latter case, one might conclude that participants will be better adjusted following their involvement in preventive programs.

ORs

Another possible index of effect in group designs is the OR which reflects the odds of a successful or desired outcome in the intervention group relative to the odds of a similar outcome in the control group. An OR can be calculated using Equations (4) or (5). An OR means that two odds are being compared, namely, the odds of success for the intervention group compared to the odds of success for the control group. The odds of success are based on the chance of success (p) divided by the chance of failure ($1-p$) for each group.

Example A. Assume a hypothetical finding indicating that 50% of the intervention group ($n=50$) graduated from high school following a mentoring program compared to 20% of the controls ($n=50$). Using Equation (4), the OR would be

$$\frac{[25\times 40]}{[25\times 10]}=\frac{1000}{250}=4.0.$$

Using proportions in Equation (5), the OR would be the same:

$$\frac{[.5/(1-.5)]}{[.2/(1-.2)]}=\frac{1.00}{.25}=4.0.$$

An OR of 1.0 indicates that the odds in each group are the same (i.e., indicating zero effect). Thus, in this case, an OR of 4.0 would reflect that the odds of graduation for the intervention group were four times higher than the odds for the controls. This OR should *not* be interpreted as indicating that four times more students in the intervention than the control group graduated; such an interpretation should be based on the calculation of relative risk ratios which are determined differently than ORs. Relative risk ratios compare the probability of an event occurring in one group compared to the probability of the same event occurring in another group. For example, the relative risk ratio for graduation in Example B is (50/100) for the intervention group divided by 20/100 for the control group, which equals 2.5. A person in the intervention group is 2.5 times more likely to graduate than someone in the control group (Liberman, 2005).

Sometimes there is confusion when authors present data for a negative or undesirable outcome. For example, assume a hypothetical result indicating that 20% of an intervention group but 50% of a control group no longer met diagnostic criteria for a clinical disorder following treatment. Using Equation (4), the odds of the intervention group having a diagnosis are 0.2/0.8 or 0.25 while the corresponding odds for the controls is 0.5/0.5 or 1.0. The OR in this case is 0.25/1.0 or 0.25. The odds for the intervention group is thus one-fourth the odds of the controls, or, alternately, the odds for the control group is four times higher than the intervention group.

The theoretical range of the OR is from zero to infinity. Computationally, if all members of the experimental group fail, the OR will be zero; if all in the control group fail, the OR cannot be estimated because the denominator would be zero. In such cases, however, it is recommended (Lipsey & Wilson, 2001) to impute 0.5 into Equation (12) for the zero cell so that the OR can be calculated [alternately, use a p-value of .99 and .01 instead of 1.00 or zero in Equation (10)]. Otherwise, a noteworthy finding in which all of the experimental or control group met criterion could not be described as an OR.

Because of its statistical properties, an OR is preferred over the SMD (Fleiss, 1994) as an index of effect in group designs when the outcome data are truly dichotomous (e.g., being arrested or not, graduating or not, meeting diagnostic criteria or not). Researchers sometimes treat continuous data dichotomously by creating groups (e.g., by using some cut-off score or doing median splits). However, others have shown that such practices should be avoided in the absence of a strong rationale and can reduce the estimate of an ES by as much as 20% (Cohen, 1983; Hunter & Schmidt, 2004). ORs are used much more frequently in medicine because of the extent to which dichotomous outcomes are studied (e.g., mortality, survival, rate of infections, presence of myocardial infarctions) but they could be used more frequently in the social sciences if more researchers were aware of them. Haddock, Rindskopf & Shadish (1998) offer a primer on methods and issues related to ORs.

ESs for Correlational Designs

Product–Moment Correlation

The index of choice in a correlational design is the product–moment correlation coefficient, r, which is calculated

in the traditional fashion, and is obtainable in the standard output of statistical packages; *r* is a widely used index of effect that conveys information both on the magnitude of the relationship between variables and its direction (Rosenthal, 1991). The possible range of *r* is well known: from −1.00 through zero (absolutely no relationship) to +1.00. Variants of *r*, such as rho, the point-biserial coefficient, and the phi coefficient can also be used as an ES.

In most cases, when multiple regression analyses are conducted, the magnitude of effect for the total regression equation is simply the multiple *R*. The unique contribution of each variable in a multiple regression can be determined by using the *t*-value that is provided by statistical packages when that variable enters the regression. Apply Equation (6) from the appendix. In structural equation modeling, standardized path coefficients assessing the direct effect can be used as *r;* there can also be indirect and total effects calculated, of which a direct effect is only a part (Kline, 2005). However, because path coefficients are similar to partial correlations, they are not the same as zero-order correlations. Rosenthal (1991) and Hunter and Schmidt (2004) provide excellent discussions of *r*-related ESs and different research circumstances affecting their calculation and application.

Aides in Calculating ESs

The basic numbers needed for calculating effects (e.g., *n*s, means, *SD*s, and *r*) are available in the standard output of the usual statistical software packages but only a few ES options are accommodated in current software programs. It is a good idea to re-check hand calculations of ESs independently in much the same way that observations and coding is checked for reliability. Meyer, McGrath and Rosenthal (2003) have prepared a helpful guide for calculating various types of ESs using SPSS and SAS syntax. Several additional resources are now available on the web that include user-friendly information regarding ESs and their interpretation, and macros for the easy calculation of effects for many common situations. Some sites also provide the means for transforming one effect into another (e.g., SMDs into *r*). The following sites developed by David Wilson, Lee Becker, and Larry Lyons and Wendy Morris are recommended (http://mason.gmu.edu/~dwilsonb/home.html, http://web.uccs.edu/lbecker/Psy590/escalc3.htm, and http://www.lyonsmorris.com/lyons/metaAnalysis/index.cfm).

Considerations Regarding the Choice of an ES

Although the effects discussed here (SMD, *r*, and OR) can be easily transformed into each other, the choice of the most appropriate effect in each situation is important. When outcomes are truly dichotomous (high school graduation, diagnostic status), the OR is usually the best choice. In group designs when continuous outcome measures are used, and raw mean differences cannot be compared across studies, SMDs are preferred, and *r* is best for correlational studies. However, one's research goals and the statistical assumptions related to the use of an ES statistic affect the choice of the most appropriate measure.

For example, McGrath and Meyer (2006) and Ruscio (2008) offer a valuable discussion of the relative merits of SMDs and *r* in two group designs when one variable is dichotomous and the other is not. These authors have shown that the superiority of one index over another depends on several factors such as unequal *n*s and heterogeneous variances in the two groups. The degree to which these factors are present will influence the calculation, value, and interpretation of an ES, and in some cases might require the use of a nonparametric alternative to SMD and *r* (Ruscio (2008).

How to Interpret ESs

Once authors calculate ESs, they need to answer this blunt but succinct question: ''What does the ES mean?'' An ES is simply a number and its meaning and importance must be explained by the researcher. An ES of *any* magnitude can mean different things depending on the research that produced it and the results of similar past studies. Therefore, it is the researcher's responsibility to discuss the importance of his or her findings and this information requires comparing current effects to those obtained in previous work in the same research area.

Judging ES in Context

Many authors still routinely refer to Cohen's (1988) comments made in reference to power analysis that SMDs of 0.20 are ''small'' in magnitude, those around 0.50 are ''medium'' and those around or above 0.80 are ''large.'' In terms of *r*, Cohen suggested corresponding figures of 0.10, 0.30, and 0.50. What many researchers do not realize is that Cohen offered these values cautiously as a general rule of thumb that *might* be followed in the absence of knowledge of the area and previous findings (Volker, 2006). Unfortunately, too many authors have applied these suggested conventions as iron-clad criteria without reference to the measurements taken, the study design, or the practical or clinical importance of the findings. Now that thousands of studies and meta-analyses have been conducted in the social sciences, Cohen's (1988)

general conventions do not automatically apply. Moreover, assuming that ''large'' effects are always more important than ''small'' or ''medium'' ones is unjustified. It is not only the magnitude of effect that is important, but also its practical or clinical value that must be considered.

For example, based on what has been achieved in many different types of interventions, educational researchers have indicated that ESs around 0.20 are of policy interest *when they are based on measures of academic achievement* (Hedges & Hedberg, 2007). This would suggest that a study with an effect of 0.20, which at first glance, might be misconstrued as a ''small'' effect if one automatically invokes Cohen's original conventions, can be an important outcome in some research areas. In other words, authors should examine prior relevant research to provide some indication of the magnitude obtained in previous studies on different types of outcomes so that current findings can be placed into an appropriate context. *Do not reflexively resort to Cohen's (1988) Conventions*.

Three guidelines are offered with respect to evaluating ESs in context. The first is to consider the source, that is, the quality of the research that produces the effect. This refers to both new findings and prior relevant research. The second guideline is to compare apples to apples, that is, make comparisons across similar research conditions, particularly when it comes to the type of outcome measure. The third is to consider the findings' clinical or practical significance.

The Quality of the Research Study Matters

ESs do not appear magically out of thin air, but are produced by specific research procedures executed within a specific research design. Methodological features often influence the magnitude of obtained effects such as the general research design, assessment methods, and the type of outcomes examined (Wilson & Lipsey, 2001). For example, based on earlier comments, the ESs achieved in *N* of 1, one-group pre-post designs, and control group designs are not directly comparable because the standards for judging the magnitude of effect are influenced by these designs. Comparing ESs from quasi-experimental and randomized designs should be considered in the light of prior research. Sometimes, randomized and quasi-experimental designs in child outcome research yield different ESs and sometimes they do not (cf. Durlak & Wells, 1997; Haney & Durlak, 1998; Wilson et al., 2001). This makes knowing the findings of prior work in one's area a must.

Different research features can easily make a lower magnitude ES achieved on one measure more important than one of higher magnitude based on another type of measure. In general, researchers place more confidence in more rigorously conducted investigations although what constitutes rigor varies from area to area. Usually, there is more confidence in a finding based on a more objective measurement strategy than one based solely on self-report. The reader can imagine other research situations where one research method is preferred over another. In other words, do not focus only on the magnitude of effects; some ESs carry more weight in the arena of scientific opinion than others because they are based on stronger research methodology.

Comparability of Outcomes is Important

With respect to the second guideline for interpreting ES appropriately, it is essential to consider the outcomes being compared. When measures are comparable across studies, the mean effect and the distribution of effects reported in meta-analyses of previous research can be used.

Importance of CIs

Finch and Cumming (this issue) have focused on the importance of calculating and correctly interpreting CIs around a sample mean in order to put new research findings into an appropriate context. The same applies to ESs. CIs are an important way to evaluate the precision of a study's findings by providing a range of likely values around the obtained ES. The equation for calculating 95% CIs around *g* (Hedges & Olkin, 1985) is Equation (3).

Example B. Suppose a *g* of 0.50 was obtained based on *n*s of 35 and 40 for an intervention and control group. Applying Equation (3), the 95% critical value is 1.96, the *SD* would be 0.23 and the CI around the mean would range from +0.04 to +0.96. This is a wide interval that is likely to include the values of most previously reported ESs in a research area. In other words, CIs represent a range of plausible values for the true population value of the ES as compared to the ES obtained in one study. Based on 95% CI, in 95 of 100 replications of the study, the true population value of the ES would fall within the CIs (i.e., between the values of +0.04 and +0.96.

Therefore, in Example B, even if all the previously reported ESs in this area were between 0.15 and 0.25, the new estimate of the ES is not very precise, and it would be inappropriate for the researcher to conclude the new ES is ''better than'' or ''significantly higher than'' prior findings. Hedges and Olkin (1985) describe how to determine if ESs differ significantly from each other based on their CIs.

Equation (3) emphasizes the influence of sample size on CIs. For example, if the group sizes in Example B were

Table I. Guidelines for Calculating, Reporting, and Interpreting ESs

1. Choose the most suitable type of effect based on the purpose, design, and outcome(s) of a research study.
2. Provide the basic essential data for the major variables[a]
   (a) for group designs, present means, standard deviations, and sample size for all groups on all outcomes at all time points of measurement
   (b) for correlational studies, provide a complete correlation matrix at all time points of measurement
   (c) for dichotomous outcomes, present the cell frequencies or proportions and the sample sizes for all groups
3. Be explicit about the type of ES that is used.
4. Present the effects for all outcomes regardless of whether or not statistically significant findings have been obtained.
5. Specify exactly how effects were calculated by giving a specific reference or providing the algebraic equation used.
6. Interpret effects in the context of other research
   (a) the best comparisons occur when the designs, types of outcomes, and methods of calculating effects are the same across studies.
   (b) evaluate the magnitude of effect based on the research context and its practical or clinical value.
   (c) if effects from previous studies are not presented, strive to calculate some using the procedures described here and in the additional references.
   (d) use Cohen's (1988) benchmarks, only if comparisons to other relevant research are impossible

[a]These data have consistently been recommended as essential information in any report, but they also can serve a useful purpose in subsequent research if readers need to make any adjustments to your calculations based on new analytic strategies or want to conduct more sophisticated analyses. For example, the data from a complete correlation matrix is needed for conducting meta-analytic mediational analyses.

doubled to 70 and 80, the 95% CIs would be substantially reduced and would range from 0.41 to 0.59. On a relative basis, this is a more precise estimate of the ES indicating the value of larger samples in research studies.

Practical or Clinical Significance is Important

In addition to the magnitude of an ES and the research context, researchers need to evaluate its practical or clinical significance. The terms, *practical significance* and *clinical significance*, are used interchangeably to reflect the extent to which there has been a meaningful change in participants' lives. It is possible to use ESs to gauge the practical significance of findings in several ways.

First of all, some raw scores that can be used to assess effects have inherent practical value. For example, the intervention group demonstrated a 40% reduction in symptoms relative to the controls following intervention. Direct comparisons using normed measures that have known and accepted clinical cut-off scores can be made. For example, a researcher might be able to say that only half of a pediatric sample had scores after intervention indicative of clinically relevant problems compared to three-quarters of the controls.

With respect to SMDs and *r*, there is no standard way to gauge practical significance and authors have different preferences. Three options are discussed here. First, the type of outcome bears examination. For example, reviews of psychosocial interventions for youth indicate that some outcomes are much easier to change than others. Higher ESs are usually attained on assessments of knowledge, attitudes and social and cognitive skills than on measures of behavior, peer sociometric status, or academic achievement (Durlak & Wells, 1997; Haney & Durlak, 1998; Wilson & Lipsey, 2007).

Frequently, the outcomes that are more difficult to change may have more practical or clinical value so that a lower ES on one outcome can be more important than a higher one on another outcome. For example, an ES of 0.20 based on behavioral measures of aggression has more clinical significance than an ES of 0.60 on attitudes toward violence. An ES of 0.25 on academic achievement (see above) has more practical value than an ES of 1.00 for rates of homework completion. It is not always possible to determine the practical benefits of different types of change, but it is worth a try in order to capture the full meaning of research findings.

A second way of assessing practical value is to report an improvement index, based on Cohen's U_3 index, which converts an effect into a percentile gain manifested by the target group (http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=4&tocId=1).

One looks up in a table the *z* score (i.e., corresponding to the obtained ES) that indicates the proportion of area under the normal curve falling under that *z*-value and interprets this area in terms of percentiles. For example, an ES of 0.50 equals a *z*-value of .50 which corresponds to the 69th percentile. If there were no effect at all, both the target group and the controls would be at the 50th percentile. An ES of 0.50 thus indicates that the average person in the intervention group is at the 69th percentile on that outcome measure (or is 19 percentiles higher than average control group member). Similarly, an ES of 0.25 (see above discussion on educational research) represents a 10% percentile gain for the intervention group over controls. While a larger gain would be more desirable, many educators would probably welcome a 10% improvement in student test scores.

Still another possibility is to use Rosenthal and Rubin's (1982) binominal ES display (BESD). The BESD uses r to calculate the relative success rates (SRs) for two groups. If SMDs are used, they are first converted to r (e.g., via the relevant formula in the appendix). The SRs for two groups are determined by the formulas:

$SR = 50\% + r/2$ (converted to a percentage) for the intervention group, and

$SR = 50\% - r/2$ (converted to a percentage) for the control group.

For example, suppose an r of 0.20 was obtained. Using the above formulae, the SRs would be 50% plus $0.20/2 = 0.10$ (10%) which equals 60% for the intervention group and $50\% - 0.20/2 = 0.10$ (10%) which equals 40% for the controls. SRs are interpreted such that if there was no effect, the SRs for both groups would be 50%. The BESD has the advantage of being able to translate a seemingly ''small'' effect into a notable difference. In this case, a ''small'' r of 0.20 translates into a 20% difference in the SRs for the intervention and control group.

The BESD is intuitively appealing for dichotomous outcomes, but presents interpretative challenges for continuous measures. How should a SR of 60% be understood when the outcome is a child's level of self-esteem? Similarly, some outcomes are easier to interpret than others when translated into an improvement index (e.g., achievement test scores falling at a particular percentile). Several good discussions of the pros and cons of different methods of assessing the practical value of ESs are available (Hill, Bloom, Black & Lipsey, 2008; Randolph & Edmonson, 2005; Valentine & Cooper, 2003).

Finally, authors must inspect prior work carefully so that they do not misinterpret one of the many types of ESs used by others as comparable to the index of effect they have used. For example, variance-accounted-for effects such as η^2 should not be compared to SMDs. Sometimes, clarity about earlier findings can be elusive and will take some investigative work. It would not be surprising if some previous reports contained an ES, but did not clarify which type of effect it was or exactly how it was calculated. Several resources distinguish among the many different types of possible effects (Cooper & Hedges, 1994; Grissom & Kim, 2005; Kline, 2004).

What if Previous ESs Are Not Reported?

Because many social scientists have not reported ESs, this information will be missing in previous publications. All is not lost, however. Authors can calculate the ESs from previous work using the methods discussed here if the basic data are available (e.g., n, mean, and SD). Even if this information is missing, ESs can be estimated in many instances using other information such as t- or F-values, probability levels, and the numbers or percentages of participants with different outcomes. Several sources provide the necessary details for correlational studies and group designs (Lipsey & Wilson, 2001; Rosenthal, 2001).

In sum, ESs become meaningful when they are judged within the context of prior relevant research and the magnitude of the effect is only one factor. These judgments should be based on the characteristics of the research studies producing the effect, the type of outcomes assessed, and the practical or clinical significance of the results.

Should Corrected ESs Be Reported?

Several procedures have been developed to correct ESs in light of factors known to influence their magnitude. For example, in addition to correcting for small sample bias which was noted earlier and is strongly recommended as standard practice, it is possible to correct for unreliability of measurement, restriction of range, and attenuation (Hunter & Schmidt, 2004; Lipsey & Wilson, 2001; Rosenthal, 1991). Comparisons to other research findings will probably have to be done using uncorrected effects if previous researchers have not corrected their reported effects.

Postscript

Examining the effects from prior research is valuable not only in a post hoc fashion when comparing new results to what has been achieved in the past, but also on an a priori basis in terms of informing the original study. ES is one important factor in determining the statistical power of analyses and many research areas are characterized by a high rate of insufficiently powered designs (Cohen, 1988, 1990; Weisz, Jenson Doss & Hawley, 2005). If previous research has consistently reported ESs of low magnitude, then the sample size needed to address a research question adequately can be anticipated. For example, if an ES of around 0.50 is expected, 64 subjects are needed in each of two groups to achieve 80% power for a two-tailed test at the .05 level. However, if prior research indicates the effect is more likely to be around 0.20, to have the same level of power would require 394 participants in each group! Having 64 per group would provide only 20% power.

In sum, an ES by itself can mean almost anything. A ''small'' ES can be important and have practical value whereas a ''large'' one might be relatively less important

or persuasive. It is up to each author to use ESs to explain the meaning and practical value of their findings. Volker (2006) offers this helpful admonition: ''The easy availability of Cohen's arbitrary guidelines should not be an excuse for us to fail to seek out and/or determine our own domain-specific standards based on empirical data and reasoned arguments'' (p. 671).

Concluding Comments

It is important for consumers and producers of research to understand what ESs are, why they are important, and how to calculate and interpret them. Guidelines regarding ESs are offered in Table I. For those who currently do not have the necessary background and knowledge, this article along with the other references cited here are intended to provide readers with the information needed for many research situations.

Conflicts of interest: None declared.

Received April 8, 2008; revisions received January 9, 2009; accepted January 10, 2009

References

Allison, D. B., & Faith, M. S. (1995). Antecedent exercise in the treatment of disruptive behavior: A meta-analytic review. *Clinical Psychology: Science and Practice, 2*, 279–303.

American Psychological Association (2001). *Publication manual of the* American Psychological Association (5th ed.). Washington, DC: Author.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T.R. Kratochwill, & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Capraro, R. M., & Capraro, M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Education and Psychological Measurement, 62*, 771–782.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hilllsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis. New York: Russell Sage Foundation.*

Durlak, J. A., Weissberg, R. P. Taylor, R. D., Dymnicki, A. B., & Schellinger, K. B. Promoting social and emotional learning enhances school success: Implications of a meta-analysis. Unpublished manuscript, *Loyola University Chicago*.

Durlak, J. A., & Wells, A.M. (1997). Primary prevention mental health programs for children and adolescents: A meta-analytic review. *American Journal of Community Psychology, 25*, 115–152.

Finch, S., & Cumming, G. (2008). Putting research in context: Understanding confidence intervals from one or more studies. *Journal of Pediatric Psychology*. Advance Access published December 18, 2008, doi:10.1093/jpepsy/jsn118.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research. A broad practical approach*. Mahwah, NJ: Erlbaum.

Haddock, C. K., Rinsdkopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3*, 339–353.

Haney, P., & Durlak, J. A. (1998). Changing self-esteem in children and adolescents: A meta-analytic review. *Journal of Clinical Child Psychology, 27*, 423–433.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hill, J. C., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology, 58*, 729–740.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Liberman, A. M. (2005). How much more likely? The implications of odds ratios for probabilities. *American Journal of Evaluation, 26*, 253–266.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

Lipsey, M. W., & Wilson D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McGrath, R. E., & Meyer, B. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11*, 386–401.

Meyer, G. J., McGrath, R. E., & Rosenthal, R. (2003). Basic effect size guide with SPSS and SAS syntax. Retrieved January 21, 2008 from http://www.tandf.co.uk/journals/resources/hipa.asp.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313–324.

Randolph, J. J., & Edmondson, R. S. (2005). Using the Binomial Effect Size Display (BESD) to present the magnitude of effect sizes to the evaluation audience. *Practical Assessment, Research and Evaluation, 10*. Retrieved January 21, 2008 from: http://pareonline.net/getvn.asp?v=10&n=14.

Rosenthal, R. (2001). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

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

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*, 19–30.

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

Thompson, B. (2006). Research synthesis: Effect sizes. In J. L. Green, G. Camilli, & Patricia B. Elmore (Eds.), *Handbook of complementary methods in educational research* (pp. 583–603). Mahwah, NJ: Erlbaum.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423–432.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret effect sizes. *Journal of Counseling Psychology, 51*, 473–481.

Valentine, J., & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse. Retrieved August 22, 2008 from http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=19&tocId=5.

Volker, M. A. (2006). Reporting effect sizes in school psychology research. *Psychology in the Schools, 43*, 653–672.

Weisz, J. R., Jenson Doss, A., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology, 56*, 337–363.

Wilkinson, L., Task Force on Statistical Inference, & APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wilson, D. B., Gottfredson, D. C., & Najaka, S. S. (2001). School-based prevention of problem behaviors: A meta-analysis. *Journal of Quantitative Criminology, 17*, 247–272.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*, 413–429.

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine, 33*(Suppl. 2), 130–143.

Appendix

Some Equations for Calculating Different Types of ESs and Transforming One Effect into Another

The following notations are used throughout the equations. *N* refers to the total sample size; *n* refers to the sample size in a particular group; *M* equals mean, the subscripts *E* and *C* refer to the intervention and control group, respectively, *SD* is the standard deviation, *r* is the product–moment correlation coefficient, *t* is the exact value of the *t*-test, and *df* equals degrees of freedom.

ESs for Group Designs

(1) Calculating Hedges' *g* from means, standard deviations and *n*s

$$g = \frac{M_E - M_C}{SD \text{ pooled}} \times \left(\frac{N-3}{N-2.25}\right) \times \sqrt{\frac{N-2}{N}}$$

$$SD \text{ pooled} = \sqrt{\frac{[(SD_E)^2(n_E - 1)] + [(SD_C)^2(n_C - 1)]}{(n_E + n_C) - 2}}$$

(2) Calculating Cohen's *d* from means and standard deviations

$$d = \frac{M_E - M_C}{\text{Sample } SD \text{ pooled}} \times \left(\frac{N - 3}{N - 2.25}\right) \times \sqrt{\frac{N - 2}{N}}$$

$$\text{where sample SD pooled} = \sqrt{\frac{[(SD_E)^2 + (SD_C)^2]}{2}}$$

(3) Calculating 95% CI for *g*:

$$CI = \pm\text{Critical value at .05} \times SD \text{ of } g$$

$$SD \text{ of } g = \sqrt{\frac{N}{n_E + n_C} + \frac{g^2}{2N}}$$

(4) Calculating an OR

$$\text{OR} = \frac{ad}{bc}$$

where *a* and *c* are the number of favorable or desired outcomes in the intervention and control groups respectively and *b* and *d* are the number of failures or undesirable outcomes in these two respective groups

(5) Alternative calculation formula for OR

$$\text{OR} = \frac{[P_E/(1 - P_E)]}{[P_C - (1 - P_C)]}$$

where *PE* is the proportional success for the Experimental group, and *PC* the proportional success for the Control group.

Effects for Correlational Designs

(6) Computing *r* from an independent *t*-test

$$r = \sqrt{\frac{t^2}{(t^2 + df)}}$$

Transforming One ES Into Another

(7) Hedges' *g* computed from *r*

$$g = \frac{r/\sqrt{1 - r^2}}{\sqrt{df(n_1 + n_2)/n_1 n_2}}$$

(8) Transforming Hedge's *g* to *r*

$$r = \sqrt{\frac{g^2 n_1 n_2}{g^2 n_1 n_2 + (n_1 + n_2)df}}$$

(9) Hedges' *g* computed from Cohen's *d*

$$g = \frac{d}{\sqrt{N/df}}$$

(10) Cohen's *d* calculated from Hedges' *g*

$$d = g\sqrt{N/df}$$

(11) Transforming Cohen's *d* to *r*

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

(12) Transforming Hedge's *g* to *r*

$$r = \sqrt{\frac{g^2 n_1 n_2}{g2n_1 n_2 + (n_1 + n_2)df}}$$