

On Effect Size

Ken Kelley

University of Notre Dame

Kristopher J. Preacher

Vanderbilt University

Accepted for publication at *Psychological Methods*:

Kelley, K. & Preacher, K. J. (in press). On effect size. *Psychological Methods*.

Author note

The authors would like to thank Thom Baguley for helpful suggestions and comments to previous versions of this article.

Correspondence to this article should be addressed to Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana 46556. E-mail: kkelley@nd.edu

Abstract

The call for researchers to report and interpret effect sizes and their corresponding confidence intervals has never been stronger. However, there is confusion in the literature on the definition of effect size and consequently the term is used inconsistently. We propose a definition for effect size, discuss three facets of effect size (dimension, measure/index, and value), outline 10 corollaries that follow from our definition, and review ideal qualities of effect sizes. Our definition of effect size is general and subsumes many existing definitions of effect size. We define effect size as *a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest*. Our definition of effect size is purposely more inclusive than the way many have defined and conceptualized effect size and it is unique with regards to linking effect size to a questions of interest. Additionally, we review some important developments in the effect size literature and discuss the importance of accompanying an effect size with an interval estimate that acknowledges the uncertainty with which the population value of the effect size has been estimated. We hope that this article will facilitate discussion and improve the practice of reporting and interpreting effect sizes.

Keywords: effect size, confidence intervals, research design, research question, reporting results.

On Effect Size

Researchers are commonly advised by methodologists, journal editors, reviewers, and professional organizations to report effect sizes and their corresponding confidence intervals as a replacement for or supplement to null hypothesis significance tests (NHSTs). Although effect size is a topic often discussed as if everyone concerned is in agreement on what an effect size is, a review of the literature leads us to conclude that there is room for improvement in the conceptualization of effect size. In fact, in the methodological literature there is a great deal of variability and ambiguity regarding the definition of effect size. However, there is also a clear push for effect sizes to be more widely reported and used as a basis for communicating results and discussing the importance of those results from research studies. Indeed, some have suggested that the use of NHSTs should be abandoned and replaced with effect sizes and confidence intervals (e.g., Schmidt, 1996). Because of the growing importance of effect sizes in research, coupled with inconsistencies in definition and conceptualization of effect sizes in the literature, we believe it will be useful to define effect size in a way that encompasses the way in which effect sizes are and can be used in research, as well as to delineate various properties of effect sizes.

Much of the impetus for this work is to provide a discussion and formalization of various aspects of the “effect size movement” (Robinson, Whittaker, Williams, & Beretvas, 2003, p. 51) that are scattered across the methodological literature. In this article we provide a broad conceptualization of effect size, review guidelines and arguments for reporting effect sizes, provide examples of existing definitions of effect size, present our own definition of effect size, show that effect size consists of multiple facets, and provide corollaries of our definition in a way that is directly applicable to applied research. We hope that this article will facilitate

discussion and improve the development, reporting, and interpretation of effect sizes so that more meaningful and cumulative knowledge can come from research in psychology and related disciplines.

Existing Guidelines for Reporting Effect Size

As recommended by Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (1999), researchers should “*always present effect sizes for primary outcomes*” (p. 599). Wilkinson and the APA Task Force go on to recommend that “interval estimates should be given for any effect sizes involving principal outcomes” (p. 599). The *Publication Manual of the American Psychological Association* (hereafter “*APA Manual*”) states in its most recent edition that NHSTs are “but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results” (2010, p. 33). The *APA Manual* goes on to say “complete reporting for all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals” (p. 33). The language in the *APA Manual* (e.g., “minimum expectations”) is much stronger than that given in the previous edition of the *APA Manual* (2001), which was not as clear at conveying the importance of reporting effect sizes (e.g., section 1.10).¹

Besides the APA, other organizations have stated officially the importance of effect sizes. The American Educational Research Association’s (AERA) *Standards for Reporting on Empirical Social Science Research in AERA Publications* states that “an index of the quantitative relation between variables” (i.e., an effect size) and “an indication of the uncertainty of that index of effect” (e.g., a confidence interval) should be included when reporting statistical results (Task Force on Reporting of Research Methods in AERA Publications, 2006, p. 10). The Society

for Industrial and Organizational Psychology (SIOP) has stated policies regarding personnel selection in its *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003). The SIOP policy on personnel selection states that “the analysis should provide information about effect sizes and the statistical significance or confidence associated with predictor-criterion relationships” (p. 19). Additionally, the National Center for Education Statistics (NCES), which is the principal statistical agency within the U.S. Department of Education (DOE), has published a set of statistical guidelines in the *NCES Statistical Standards* (2002). The *NCES Statistical Standards* (2002) is “intended for use by NCES staff and contractors to guide them in their data collection, analysis, and dissemination activities” (p. 1). The *NCES Statistical Standards* states that “when the results of an analysis are statistically significant, it is useful to consider the substantive interpretation of the size of the effect ... for this purpose, the observed difference can be converted into an effect size to allow the interpretation of the size of the difference” (Section 5-1).

Editors of several journals in psychology and related disciplines have made explicit their commitment to requiring effect sizes to be reported. For example, *Educational and Psychological Measurement*, an early adopter of strong language regarding effect sizes, states “authors reporting statistical significance will be *required* to report and interpret effect sizes (Thompson, 1994, p. 845, italics in original). *Educational and Psychological Measurement* is largely a methodological journal, so it would be understandable if its policies differed from those of more substantively oriented journals. However, many substantive journals have similar policies. For example, a *Journal of Applied Psychology* editorial states that “if an author decides not to present an effect size estimate along with the outcome of a significance test, [the editor] will ask the author to provide specific justification for why effect sizes are not reported. So far,

[the editor has] not heard a good argument against presenting effect sizes” (Murphy, 1997, p. 4). Similarly, a *Journal of Consulting and Clinical Psychology* editorial states that, “evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance” (Kendall, 1997, p. 3). The author guidelines for *Psychological Science* regarding statistics begins “effect sizes should accompany major results” and does not mention NHSTs (Association for Psychological Science, 2011). Many more journals have stated as policy their strong preference or requirement for effect sizes to be reported (Vacha-Haase & Thompson, 2004 note that, at the time, 23 journals had such policies; a partial list is provided in Fidler & Thompson, 2001).

Issues of effect size reach beyond psychology, education, and management. The International Committee of Medical Journal Editors’ (ICMJE) *Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication* illustrates biomedical editors’ preference for the use of effect sizes and confidence intervals: “when possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals) ... avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important information about effect size” (ICMJE, 2007, Section IV.A.6.c). Additionally, the Consolidated Standard of Reporting Trials (CONSORT; Schulz et al., 2010, items 17a and 17b) and the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND; Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004, item 17) both state that effect sizes and confidence intervals should be reported for primary and secondary outcomes. Clearly, there is growing recognition of the importance of

reporting effect sizes along with confidence intervals for interpreting and communicating the magnitude of an effect and discussing its importance.

Existing Definitions of Effect Size

Even with the importance of effect size in modern research, there are different and conflicting definitions of effect size in the literature. Nakagawa and Cuthill (2007) discuss how effect size can mean (a) “a statistic which estimates the magnitude of an effect” (e.g., r), (b) “the actual values calculated from certain effect statistics” (e.g., $r = .3$), or (c) “a relevant interpretation of an estimated magnitude of an effect from the effect statistics” (e.g., “medium”) (p. 593). Because effect size is used in three entirely different ways, it can be problematic to know precisely what is meant by the term “effect size.” Later, we provide terms to clearly distinguish the first two uses of “effect size” documented by Nakagawa and Cuthill (2007). The third way Nakagawa and Cuthill document that the term “effect size” is used, namely as a qualitative interpretation of a quantitative value, is generally problematic and is an issue we discuss later. Notice that “effect” is used in each of the three meanings of effect size discussed by Nakagawa and Cuthill, which is an illustration of the difficulty of discussing effect sizes in a way separated from the word “effect.” We will not define effect size with the term “effect” in the definition, or “size,” so as to avoid defining effect size with one of its two root words.

Effect size is often linked to the idea of substantive significance (e.g., practical, clinical, medical, or managerial importance), which can be understood to be the degree to which interested parties (e.g., scientists, practitioners, politicians, managers, consumers, decision makers, the public at large, etc.) would consider a finding important and worthy of attention and possibly action. *Substantive significance* is context specific and can mean different things to different parties in different situations. The idea of substantive significance is more subjective

than *statistical significance*, which has a highly objective meaning (i.e., obtaining a p -value less than the specified Type I error rate).² Although objective, Aiken reminds us that even a small p -value need not imply substantive significance (1994, p. 857).

Contrary to conventional wisdom, what may seem like a trivial effect size can translate into a substantively important finding. Rosenthal and Rubin (1982) and Abelson (1985) provide illustrations of how what some may regard as a trivial effect size can translate into substantive significance. Consider the Physicians' Health Study—the Data Monitoring Board recommended early termination of this randomized study because a “statistically extreme beneficial effect on nonfatal and fatal myocardial infarction had been found” (Steering Committee of the Physicians' Health Study Research Group, 1988, p. 262). This “statistically extreme” effect can be translated into a product moment correlation coefficient of the dichotomous explanatory variable (group membership) and outcome (myocardial infarction or not) of .0337, which, when squared to obtain the proportion of explained variance, is only .0011 (e.g., Rosenthal, 1994). However, another way to interpret the results is that the physicians not assigned to the aspirin group had a 1.82 times higher odds of having a heart attack (i.e., $[189/11034] / [104/11037] = 1.82$; Steering Committee of the Physicians' Health Study Research Group, 1988). This is, by all accounts, an important finding with much practical importance.

Kirk (1996) reviews the idea of substantive significance and uses the term “effect magnitude” for supplementary measures to accompany NHSTs, in which “measures of effect size” was a special case of effect magnitude relating specifically to standardized mean differences (pp. 748-749). Kazis, Anderson, and Meenan describe effect size as “a standardized measure of change in a group or a difference in changes between two groups” (1989, p. S180). Olejnik and Algina (2003) define an effect size measure as “a standardized index” that

“estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables” (p. 434). One of the ways in which *NCES Statistical Standards* (2002) defines effect size is “standardized magnitude of the effect” (p. 131). We believe that these definitions of effect size are too narrow, as they exclude many measures that we believe are effect sizes. However, Kirk’s “effect magnitude” is more encompassing and is consistent with our broad conceptualization of effect size, which we discuss momentarily. Like Grissom and Kim, who note “effect size ... is not synonymous with practical significance,” we believe that “knowledge of a result’s effect size can inform a judgment about practical significance” (2005, p. 4).

Effect size has often been defined relative to the value of the null hypothesis for a corresponding null hypothesis significance test (NHST; Berry & Mielke, 2002; Henson, 2006; Kirk, 1996, 2002). Grissom and Kim (2005) state that “we use the label effect size for measures of the degree to which results differ from what is implied for them by a typical null hypothesis” (p. 4). An updated edition of the Grissom and Kim work states that, “whereas a test of statistical significance is traditionally used to provide evidence (attained p level) that a null hypothesis is wrong, an effect size (ES) measures the degree to which such a null hypothesis is wrong (if it is wrong) (2012, p. 5). The second way that the *NCES Statistical Standards* (2002) defines effect size is the “the departure from the null hypothesis” (p. 131). Similarly, one of the ways that Cohen (1988) defined effect size is “the degree to which the null hypothesis is false.” Vachon-Haase and Thompson (2004) define effect size as a “statistic that quantifies the degree to which sample results diverge from the expectations ... specified in the null hypothesis” (p. 473). Similarly, Thompson states that “effect sizes quantify by *how much* sample results diverge from the null hypothesis (2004, p. 608) and Creswell (2008) states “the calculation of effect size varies

for different statistical tests,” (p. 167), both of which link effect size to a NHST. We believe linking the definition of effect size to NHSTs is something that should be avoided, as effect sizes and NHSTs represent fundamentally different ways of using data to make inferences.

Furthermore, for the same set of data, the effect size calculated would be dependent on the null value specified by a researcher. For example, if a null value of the population correlation coefficient were specified at a value of .10 by one researcher and 0.00 for another researcher, for an obtained value of the correlation coefficient of .20, one researcher’s effect size is .10 whereas the other’s is .20. Defining an effect size relative to the null hypothesis specified in a NHST suffers from a fundamental problem in that each researcher could report a different effect size for the same phenomenon in the same data set.

As noted, one way that Cohen defined effect size was tied to a NHST. Another way that Cohen (1988) defined effect size was the “degree to which the phenomenon is present in the population” (pp. 9-10). Conversely, Vacha-Haase and Thompson (2004) confine their definition of effect size to “sample results” (p. 473). We see wedding the definition of effect size to either a population or a sample is overly limiting, because generally it is useful to conceptualize a population effect size as well as an effect size in a sample (i.e., both sample values and population values of effect sizes exist).

Sometimes effect size is defined specifically with respect to an independent or dependent variable, and sometimes both. For example, Miller (2007) describes an effect size as “a measure of how strong the relation is between the independent and dependent variable” (p. 147) and Vaske, Gliner, and Morgan define effect size as “the strength of the relationship between the independent variable and the dependent variable” (2002, p. 290). Similarly, Peyton’s (2005)

definition of effect size is “the magnitude of the impact of treatment on an outcome measure” (p. 186). We believe these definitions are useful but unnecessarily narrow.

Rather than providing a formal definition of effect size, some works use a definition-by-example approach (e.g., Ellis, 2010, chapter 1; Rosenthal, 1994). We believe that the definition of a term by providing examples is generally not the ideal way to articulate the full scope of the term. Although such an approach may provide representative examples, there may be ambiguity when attempting to generalize beyond the specific examples given. We believe that definition-by-example approaches to defining effect size fail to convey the rich variety of types of effect sizes that exist.

There is a wide range of restrictive definitions of effect size in the literature. We see the multitude of restrictive definitions as a hindrance to the increased use of effect sizes. Although a great deal of emphasis has been placed on using effect sizes to facilitate interpretation of the results of empirical research, many researchers rely on the results of NHSTs to come to research conclusions. Due to the limitations, inconsistencies and various issues outlined thus far, we believe the time is right to offer a general definition of effect size and delineate various properties of effect size. We hope our discussion will provide a general framework for discussions of effect sizes, assist in understanding the importance of linking questions of interest to effect sizes, and help researchers better communicate the magnitude of findings in rigorous and useful ways, which we believe will ultimately facilitate a more cumulative science.

A Definition of Effect Size

Before defining effect size, we first define “effect” and “size” in the present context. In a research context, we define *effect* as *a quantitative reflection of a phenomenon* and *size* as *the magnitude of something*. Using effect and size as core components, with consideration of how

research is reported and interpreted, leads to our definition: *effect size* is defined as *a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest*. Our definition of effect size is more than the combination of *effect* and *size*, as our definition depends explicitly on addressing a research question of interest. The question of interest might refer to central tendency, variability, association, difference, odds, rate, duration, discrepancy, proportionality, superiority, or degree of fit or misfit, among others (cf., Preacher & Kelley, 2011). Our definition of effect size is intentionally more inclusive than the ways in which many others have previously defined or conceptualized effect size.

We believe that effect size is necessarily tied to a question of interest and we regard this aspect of our definition very important. An estimated effect size is a statistic, whereas a population effect size is a parameter. However, a statistic or parameter is not necessarily an effect size. One might then ask, “what is the difference between an effect size and a statistic (or parameter)?” We regard the difference between a statistic (or parameter) and an effect size to be a function of its purpose: *an effect size is a statistic (or parameter) with a purpose, which is to quantify some phenomenon that addresses a question of interest*.

In some cases what will be classified as an effect size in one study may not be classified as an effect size in another study. The idea of an effect size being a statistic (or parameter) with a purpose is the reason why reporting effect size is so often suggested or even required when discussing the results of a study. That is, effect sizes are used for the purpose of conveying the magnitude of some phenomenon, where that phenomenon is explicitly linked to some research question. For example, in some contexts, a regression coefficient’s only purpose is to make an objective prediction of a criterion variable from a set of regressors in an automated fashion. In such situations the particular value of a regression coefficient may not be of interest – rather it is

simply used in an automated prediction equation. However, in other situations a regression coefficient is used to quantify the linear relation between a regressor and a criterion variable when holding other regressors constant in an effort to convey the strength of the relationship. The former example need not be regarded as an effect size because it is not literally addressing a question of interest (rather, it is used for automated prediction), but the latter use is explicitly addressing a question of interest (i.e., the strength of the relationship). When used as part of an automated prediction, for example, its value may never be of interest in and of itself—it is simply treated as a part of an algorithmic method of reaching a prediction. This is a completely different use than when the size of the regression coefficient is explicitly of interest to address a research question.

What about a proportion from a random sample—is it an effect size or “just” a statistic? The answer from the previous discussion implies that it depends on its purpose and on whether or not it addresses a research question. Consider the description of a particular firm, where the proportion of women in managerial roles can be regarded simply as a description of what is true: “Women hold 12% of the managerial roles at the firm.” However, the proportion of women in managerial roles can also be regarded as an effect size, as it represents the magnitude of some phenomenon for a question of interest. This effect size could be used as a comparison to the proportion of women in managerial positions in other firms of similar scope, as an impetus to better investigate why it is such a small proportion, or as a reflection of a collection of factors inhibiting women from promotion to managerial roles, among other uses. Thus, even a statistic as simple as a proportion can be used not only as a description of what exists in a data set, but to address a particular research question. Many other examples could be provided, but the idea is that an effect sizes addresses a question of interest.

Notice that our definition is not wedded to any particular null hypothesis or NHST. In fact, we believe that linking the definition of effect size to a null hypothesis or NHST should be avoided because effect size and null hypotheses represent two fundamentally different ways of using data. Although the move to a more effect size based literature, it could be argued, is motivated by a move away from NHSTs (Robinson et al., 2003, p. 51), our purpose is not to disparage null hypothesis significance testing or the framework in general. Rather, we believe it is important to note the conceptual independence of effect sizes and NHSTs. However, if a connection must be made between effect sizes and NHSTs, it is that NHSTs are dependent on effect size, not the reverse, as NHSTs involve testing whether a population effect size differs from some stated null value. Effect size itself need not invoke a null hypothesis or a NHST in order for it to be used as a way to express the size of an effect, but it should always be accompanied with a confidence interval to explicitly show the uncertainty associated with the estimate.

Regarding definitions that link effect size to the dependent variable, sometimes there is no clear distinction between independent and dependent variables. For example, the correlation coefficient between two variables can be of interest even if one variable is not treated as a predictor of the other (e.g., the correlation between two independent variables in the context of multiple regression). In such a situation the correlation may be of interest to assess collinearity, which can affect statistical power and accuracy of parameter estimation. The correlation coefficient also extends to a multivariate effect size in canonical correlation analysis, where there may be several variables in each of two sets, and linear composites of the sets are correlated. Also, for evaluating fit or misfit in a structural equation modeling context, often the discrepancy

between observed and model-implied covariances for an entire model is evaluated, and it is difficult to conceptualize this situation in independent-dependent variable terms.

Although it is more general than many definitions of effect size, there is support for our broadly conceptualized definition of effect size. For example, Fidler, Thomason, Cumming, Finch, and Leeman (2004) review the ways in which confidence intervals and effect sizes are used and reported in research. In a section titled “Effect Sizes,” Fidler et al. (2004) note the type of effect sizes they coded as part of their study. They state that “we coded any effect size measures” and go on to list more than 12 specific types of effect size measures, including means, odds ratios, percentages, proportions, and accounted for (or explained) variance statistics. Additionally, as noted previously, one way that Cohen defined effect size was the “degree to which the phenomenon is present in the population.” Although Cohen’s definition may implicitly invoke a null hypothesis of zero, the generality of the definition speaks to a broad conceptualization of effect size, albeit only in the context of the population.

The Facets of Effect Size

Within the context of effect size, we have identified three facets that are important to delineate. By facet, we mean a particular aspect of effect size that corresponds to how the term is used. Similar to how Nakagawa and Cuthill (2007) discuss three meanings of effect size, we try to formalize three “facets” to how we conceptualize effect size. The first facet of effect size addresses the type of information of interest; the second facet is the operationalization of the effect size via the equation that links the data, statistic, or parameters to the effect size; and the third facet is the particular value of the effect size. More formally, the facets are *effect size dimension*, *effect size measure/index*, and *effect size value*, each of which is defined below, with examples illustrating the meaning of each of the facets.

Facet 1: Effect Size Dimension

In physics, dimensions are regarded as generalized units (Ipsen, 1960, section 3.6; Carman, 1969; Ellis, 1966, p. 142). The basic idea of a dimension is that it identifies an abstraction of the variable of interest, but not the units with which the abstraction will be measured. Example dimensions in physics are length, weight, density, force, and energy. A dimension can be operationalized in different units (e.g., the dimension of distance can be operationalized in units of millimeters, inches, miles, or light-years, among others).

Effect size dimension is an abstraction of a quantifiable quality in a generalized way that does not have a particular unit. An example of an effect size dimension is variability, which can be operationalized in units of variance, standard deviation, range, interquartile range, et cetera. Notice that the effect size dimension of variability is itself not a specific unit, rather, it is an abstraction related to the degree to which values differ – variability is thus a quality that will be quantified in some way. Momentarily we discuss effect size measure, which is the way in which the effect size dimension is operationalized in a particular context.

Consider effect size dimension in an applied context. Suppose interest concerns the relationship between two variables. The dimension of relatedness (or some similar term that describes the dimension) can be operationalized in the form of a correlation coefficient, covariance, regression coefficient, among others. The dimension of relatedness is the abstract concept and not any particular operationalization of the dimension. That is, the effect size dimension provides the general idea (i.e., abstraction) of the way in which the question will be addressed. Momentarily (in corollary 3 in the next section) we discuss how dimensions are formalized in the effect size context.

Facet 2: Effect Size Measure

Effect size measure, or synonymously *effect size index*, is a named expression that maps data, statistics, or parameters onto a quantity that represents the magnitude of some phenomenon. That is, it is the equation that defines the particular implementation of the effect size dimension or dimensions of interest. For example, an effect size measure used to operationalize the effect size dimension of separation between the treatment group and the control group means is the standardized mean difference, which is defined as $d = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}$, where \bar{X}_j denotes the mean of the j th group ($j = 1, 2$) and s_{pooled} is the square root of the unbiased estimate of the within group variance (i.e., the square root of the mean square error). That is, the effect size measure is the above noted equation that maps the statistics onto the particular effect size.

Another example of an effect size measure is the root mean square error of approximation (RMSEA or ε), which is an operationalization of the effect size dimension of model fit, specifically in the context of structural equation models (other such effect size measures are the normed fit index, the goodness of fit index, adjusted goodness of fit index, and Tucker-Lewis index, among others). The effect size measure of the RMSEA is defined as

$$\varepsilon = \sqrt{\max\left\{0, \frac{\hat{F}_0}{\nu}\right\}},$$

where \hat{F}_0 is the estimate of the population maximum likelihood discrepancy function and ν is the number of degrees of freedom. Each such effect size measure has its own implementation of model misfit. The idea of the effect size index is that it has a very precise way in which the data, statistics, or parameters are used to implement a particular effect size dimension in order to address some question of interest.

When an effect size measure is applied to data, statistics, or parameters, a real number results that we term the *effect size value*. This value is the realization of a particular effect size measure, that itself was a particular implementation of an effect size dimension. The effect size value is literally the magnitude of some phenomenon as discerned from the data, statistics, or parameters. For example, the effect size value obtained for the effect size measure “the standardized difference between means” that operationalized the separation between the treatment group and the control group means in a particular study may be $d = .70$. The value of $.70$ is a real number that results from applying data, statistics, or parameters to an effect size measure (i.e., an equation or algorithm) that operationalizes a particular effect size dimension (i.e., a generalized abstraction). Another example of an effect size value is the value observed for the misfit between the theoretical model and a data set, such as $\hat{\epsilon} = .0375$.

Each of the three facets of effect size (i.e., effect size dimension, effect size measure, and effect size value) will be referred to simply as “effect size” in different contexts. This generally poses no problem in practice, as the context in which the term “effect size” is used will usually provide clarification to determine the particular facet (i.e., if it refers to effect size dimension, an effect size measure, or effect size value) being referenced. However, for any effect size value to make sense, the specific effect size measure necessarily needs to be clearly stated. Stating an effect size dimension and then providing an effect size value does not convey how the effect size dimension was implemented, that is, the particular effect size measure is unknown (or at least ambiguous). For example, stating that “the relationship between X_1 and Y is $.7$ ” is not sufficient, because the $.7$ value could be an unstandardized regression coefficient, a standardized regression coefficient, a regression coefficient where only the regressors (and not Y) are standardized, or some type of correlation coefficient (e.g., partial or otherwise). As is discernible from our

definition of effect size dimension, it is a more general term that speaks to the type of effect size that is of interest, whereas effect size measure is very specific regarding the way in which effect size is operationalized. Effect size value is then a real number from an expressly stated effect size measure that conveys information about some type of effect size dimension of interest based on data, statistics, or parameters.

Corollaries from the Effect Size Definition

In this section we discuss 10 corollaries that stem from our definition of effect size, which allows us to clarify various aspects and types of effect sizes. We discuss these corollaries because we believe they will help provide clarity to researchers. However, the failure to discuss a topic as a corollary does not imply that the particular aspect is not a corollary or is in some way unimportant. For each of the corollaries, we provide an overview and examples.

Corollaries

1. Effect sizes can represent sample values, population values, or theoretical values.

An effect size obtained from data is a sample value, which is used to estimate the corresponding population value. In many cases the population value is a theoretical value that is unknowable. However, in other cases the population value can be obtainable from perfectly reliable census data. How often “perfectly reliable census data” are available in a particular context, or even if they can be, is not relevant for the corollary. Rather, the idea of the true population value of some effect size existing is the point we want to emphasize. Additionally, theoretical values that may or may not be equal to the sample or population values also exist. For example, a theory might predict that the true correlation coefficient between variables X and Y is 0.00. The population value may in fact be .10 and the sample value may be .175. The fact that there are three different effect size values (i.e., sample, population, and theoretical) poses no

particular problems. Of course, whenever the situation is not abundantly clear from the context, the particular type of effect size value (sample, population, or theoretical) should be clearly noted.

2. *Effect sizes can quantify absolutely or comparatively.*

Some effect sizes are absolute in nature, in the sense that the effect size does not require some referent value for its interpretation. Other effect sizes, however, are comparative in nature, in which the interpretation of the effect size does require some referent value or values. The referent values can be from another group, situation, model, or theoretically interesting benchmark, such as the maximum, minimum, or baseline value, among others. Comparative effect sizes are used when there is some explicit comparison that is important to quantify. For example, the number of participants that relapse after taking part in a smoking cessation program is an absolute (referent free) effect size measure. However, the number of participants that relapse after taking part in the smoking cessation program *relative* to the total number of participants (i.e., the proportion of participants that relapse) is a comparative effect size. For another example of a comparative effect size, consider the number needed to treat (NNT) in the context of two groups with binary outcomes (e.g., success or failure). The NNT is “the estimated number of patients who need to be treated with the new treatment rather than the standard treatment for one additional patient to benefit” (Altman, 1998, p. 317). The number needed to treat is defined as

$$NNT = 100 \left(\frac{1}{p_N - p_S} \right),$$

where p_N and p_S are the proportions of participants with successful outcomes in the new and standard treatments, respectively (Altman, 1998).

3. *Effect sizes can be dimensionless, unidimensional, or multidimensional.*

Recall that the first facet of effect size is effect size dimension, which is a generalized abstraction of a unit. Determining the number of dimensions of a particular effect size measure is the basis of this corollary. In physics a dimensional analysis “treats the general forms of equations that describe natural phenomena” (Langhaar, 1951, p. v), so that a variable can be decomposed into its basic dimensions. Another way to conceptualize a dimensional analysis is as “a method by which the variables which characterize a phenomenon may be related” (Ford & Cullmann, 1959, p. 11). In a dimensional analysis, a functional relationship is given that “remains true no matter what the size of the units in terms of which the quantities are measured” (Bridgman, 1922, p. 1). For example, consider the following dimensions, where brackets are used to denote dimension and $\overset{d}{=}$ means is dimensionally equal,

$$[Time] \overset{d}{=} [Time];$$

$$[Length] \overset{d}{=} [Length].$$

The dimensions of $[Time]$ and $[Length]$ are each unidimensional because they do not depend on other abstractions, which is why the left-hand and right-hand sides of the equation are the same, as they cannot be further reduced. When a dimension does not depend on other dimensions, that is, when it is unidimensional and cannot be reduced—it is a fundamental dimension considered to be a basic building block. However, dimensions such as $[Velocity]$ are multidimensional because they are defined in terms of other dimensions:

$$[Velocity] \overset{d}{=} \frac{[Length]}{[Time]}.$$

Notice here that the left-hand and right-hand sides differ, unlike the case for $[Time]$ and $[Length]$ above, because a multidimensional abstraction (on the left-hand side) is represented as a function

of its unidimensional components (on the right-hand side). Multidimensional phenomena, such as [*Velocity*], are “derived magnitudes” that are obtained in a more complicated way than by comparing the phenomena to a particular unit (e.g., as both [*Length*] and [*Time*] are each obtained in the dimensional equation above).

We extend the idea of a dimensional analysis to an effect size context, where an effect size measure is decomposed into its basic dimensions. Dimensional analysis in the effect size context is useful because it allows the number of dimensions of an effect size measure to be determined. As we show, effect size measures can be multidimensional, unidimensional, or dimensionless, which impacts the way in which an effect size can be interpreted. We expect dimensional analyses to be most relevant to methodologists studying various effect sizes that may be appropriate in a particular context. However, applied researchers will also benefit by better understanding the dimensions that combine to form various effect sizes.

Take, for example, the population effect size measure of the mean difference, defined as

$$\Delta = \mu_1 - \mu_2$$

in the population, where μ_j is the population mean for the j th group ($j=1,2$). The effect size value of a mean difference is based on only a single dimension, namely central tendency.

Although there are two instances of central tendency in the effect size measure of Δ , there is only one dimension (because those two instances of central tendency are in fact the same dimension). More formally,

$$[\delta] = [Central\ Tendency]^d - [Central\ Tendency]$$

and thus consists of only a single dimension, central tendency. The fact that a difference is taken between two measures of the same dimension is not problematic, as dimensional analysis uses dimensional equations and not ordinary algebraic equations, in which

$[Central\ Tendency] - [Central\ Tendency]$ would be 0. Recall that dimensional analysis describes a phenomenon in terms of the mathematical form of the basic dimensions of the phenomenon. For δ , the effect size measure used to operationalize $[Central\ Tendency]$ here is the mean, but any other measure of $[Central\ Tendency]$ (e.g., the median) could have been used without the dimension itself changing, just the operationalization of the dimension via a different effect size measure (e.g., the difference between medians).

An example of an effect size measure that is multidimensional is the standardized mean difference, which is defined for the population as

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

where σ is the common within group population standard deviation. The effect size δ consists of dimensions of $[Central\ Tendency]$, operationalized as the mean (of which a difference is taken), and of $[Variability]$, operationalized as the common standard deviation:

$$[\delta] = \frac{d [Central\ Tendency] - [Central\ Tendency]}{[Variability]}. \text{ Thus, the standardized mean difference is a}$$

bidimensional effect size because it is composed of two fundamental dimensions (they are fundamental because they are at the most basic level).

Some effect size measures are dimensionless abstractions, which is a term used when a dimensional analysis reveals that the effect size is based on no specific dimensional component (notationally as $[Effect\ Size\ Measure] = [1]^d$). Note that saying an effect size has dimension of $[1]$ is different than saying it has a single dimension, such as $[Variability]$. Dimensionless variables describe what can be considered a *natural variable*, which when realized is a *natural number* that does not depend on any arbitrary convention of expression (such as a particular scaling of an abstraction, e.g., kilograms). Natural numbers have a natural unit of 1. For a concrete example

from physics, consider the *slenderness* of a cylinder, defined as the ratio of height to diameter, both of which are measures of distance (i.e., the ratio of distance along the cylinder to the distance around the cylinder). A dimensional analysis shows, then, that slenderness is a dimensionless variable that produces a dimensionless number, as it is a ratio of two distances ($[Slenderness]=[Distance]/[Distance]=[1]$). Regardless of the particular scaling of distance used to measure the height and diameter, slenderness remains the same because the same dimensions are in the numerator and denominator and are thus “divided out.”

Consider the estimated squared multiple correlation coefficient (i.e., the estimated coefficient of determination), which is defined as the ratio of the sums of squares due to the regression model to the total sums of squares,

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2},$$

where Y_i is the score of the dependent variable for the i th individual ($i = 1, \dots, N$), \bar{Y} is the mean of the dependent variable, and \hat{Y}_i is the model implied value from the regression model for the i th individual. For R^2 , the effect size dimension in the numerator (i.e., [*Sum of Squared Deviations*]) is the same as the effect size dimension in the denominator (i.e., [*Sum of Squared Deviations*]). Thus, R^2 is a dimensionless effect size (i.e., $[R^2]=[1]$). Again, for clarity, we are not saying that R^2 is an effect size with 1 dimension, but rather that its dimensionless and its dimensional components, which there are none, is represented as [1], due to the dimensions of sum of squares being divided out. The practical utility of a dimensional analysis comes from having a better understanding of how dimensions affect the interpretation of an effect size.

4. *Effect sizes can be standardized, partially standardized, or unstandardized.*

An unstandardized effect size is one whose interpretation is dependent on the units of the measurement instrument. Examples of unstandardized effect sizes are (a) unstandardized regression coefficients based on the observed data from one or more different measurement scales, (b) mean differences based on untransformed (i.e., raw) data, (c) path coefficients in a structural equation model based on the covariance matrix (not the correlation matrix), and (d) the product of unstandardized regression coefficients (or path coefficients; e.g., the indirect effect in mediation models).

In addition to or instead of a dimensional analysis, what we term a *unit of measurement analysis* can also be performed in a manner analogous to the dimensional analysis. The idea is to determine if the measurement units cancel, creating a type of effect size that needs no unit label, termed a standardized effect size. This is fundamentally different than a dimensional analysis but may initially seem similar. The difference concerns whether the analysis concerns the dimensions or the units of measurement. A standardized effect size is one in which the measurement units themselves cancel, not necessarily but possibly the dimensions, too, so that the particular units of the measurements are no longer wedded to the interpretation of the effect size. Note that a standardized effect size can be dimensionless or have some number of dimensions. Examples of standardized effect sizes are (a) standardized regression coefficients (e.g., when regression analysis is based on *z*-scores of the outcome variable and all regressors), (b) the standardized mean difference, (c) the coefficient of variation, and (d) the standardized solution from a structural equation model. For a standardized effect size, knowledge of the particular units of the measurement scale are not needed in order to interpret the value of the effect size. In this sense, a standardized effect size is similar to a dimensionless effect size, but the distinction between the two is important because they represent two different concepts.

Consider again the population standardized mean difference. The numerator is in terms of whatever units of measurement represent the phenomenon of interest (e.g., reaction time), but so too is the denominator. Thus, the measurement units expressing the magnitude in the numerator and the units expressing the magnitude in the denominator cancel, leading to an effect size measure that needs no unit label due to the measurement units canceling.

Another type of scaling of effect sizes is partial standardization. For example, the standardized solution in a multiple regression model is one in which the dependent variable and all independent variables are all standardized. However, standardizing the independent variables but not the dependent variable leads to regression coefficients that are partially standardized. The interpretation of such coefficients is that each coefficient is interpreted to mean the expected change in the unstandardized criterion variable for a 1 standard deviation change in the standardized regressor variable, holding other standardized regressors constant. More generally, an effect size is said to be partially standardized when at least one component of the effect size is standardized, but at least one component unstandardized. In such situations, the interpretation of the effect size is at least partially based on one or more specific measurement units.

Although dimensionless effect size measures from Corollary 3 and standardized effect size measures from this Corollary (Corollary 4) may seem similar, they are theoretically distinct. Recall that a dimensionless quantity is one that has dimension of [1]; that is, it has no dimensions (e.g., due to the same dimensions in the numerator and denominator canceling). An effect size value can have multiple dimensions and (a) need no specific measurement unit label (e.g., the standardized mean difference is not wedded to any particular measurement unit, such as scores on a depression inventory, and is invariant to linear transformations) or (b) be wedded to specific measurement units (e.g., an unstandardized regression coefficient, which is scaled in terms of the

particular measurement scales of the regressor and outcome variables). An effect size can also have a single dimension and (a) need no specific measurement unit label (e.g., the cardinal numbers used to measure the size of a set, which conveys the same information regardless of the label of the units contained within the set) or (b) have a single dimension and be wedded to a specific measurement unit (e.g., mean difference on a performance scale depends on the scale of the variable in which a mean difference was taken). Additionally, an effect size can be dimensionless and (a) need no specific measurement unit label (e.g., the squared multiple correlation coefficient is not wedded to the measurement scales of any variable in the model) or (b) be specific to a particular unit of measurement (e.g., the ratio of unstandardized regression coefficients involving regressors of different scales).

5. Effect sizes can be base-rate dependent or base-rate independent.

The idea of an effect size being base-rate dependent versus base-rate independent is that a base-rate dependent effect size is fundamentally linked with the proportionality of the size of the sample or population to which the effect size corresponds. Consider the point biserial correlation coefficient, which is defined in part by the proportion of a sample or population accounted for by each of two groups. In particular, when there are two groups and a continuous outcome variable, the sample point biserial correlation can be written as

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_2}{S_X} \sqrt{p_1 p_2},$$

where p_1 is the proportion in the first group ($p_2 = 1 - p_1$ is thus the proportion in the second group) and S_X is the standard deviation of the data (i.e., using the grand mean) with N (not $N-1$)

as the divisor (i.e., $S_X = \sqrt{\sum_{i=1}^N (X_i - \bar{X}_{..})^2 / N}$, where $\bar{X}_{..}$ is the grand mean of the $N X_i$ scores;

Cohen, Cohen, West, & Aiken, 2003, section 2.3.3). Whereas the point biserial correlation

depends in part on the size of the groups via their proportions, the standardized mean difference is independent of group size, making it a base-rate independent effect size. Such a distinction can be important for interpretation purposes. McGrath and Meyer (2006) discuss how the point biserial correlation coefficient and the standardized mean difference can lead to different conclusions when applied to the same data when one variable is a grouping variable and the other is a continuous variable.

The point biserial correlation coefficient is not the only base-rate dependent effect size, of course. Consider again the example of the proportion of women in managerial roles at a particular organization. The actual count of women in managerial roles is base-rate independent. However, the proportion of women in managerial roles depends not only on the count of women managers, but also the total count of managers. This proportion is the base-rate itself of women managers. Odds are also base-rate dependent, because they compare the number of “successes” with the number of “failures.” Likewise, odds ratios are the ratio of odds from two different groups and are themselves based on the proportionality of successes to failures for two groups. The point is, when an effect size depends on a base-rate – that is, a proportion in a population or sample – the effect size is base-rate dependent.

6. Effects sizes can quantify phenomena that are causal or observational.

Effect sizes may apply in causal situations or in observational situations; there is no special difference between causal and observational settings with regards to the effect size itself, as the same effect sizes can be generally be used when quantifying causal relationships or associations. A major difference between effect sizes that quantify causal versus observational, however, is how the effect size is communicated and interpreted. For example, consider a simple linear regression in which the unstandardized regression coefficient is 5. It might be tempting to say “a

1 unit change in the regressor *causes* a 5 unit increase in the criterion variable.” However, such language would be appropriate only if the values of the regressor were randomly assigned to a random sample from a population of interest, the independent variable preceded the dependent variable with sufficient time to allow the causal process to manifest, alternative explanations for the effect can be dismissed, and the assumption of linear relationships among variables was satisfied. Certainly one might believe an observational relationship is causal due to theoretical arguments, especially after “controlling” for various other variables thought to be correlated with both the regressor or the independent variable, but there may well be one or more lurking variables that actually influence the change in the outcome variables. In short, effect sizes in and of themselves are distinct from issues of causality. That is, inferring causality adds requirements that do not influence the use of effect sizes (e.g., Pearl, 2009).

7. Effects sizes can quantify phenomena that are omnibus, targeted, or semi-targeted/semi-omnibus.

An omnibus effect size is one that quantifies an overarching effect, whereas a targeted effect size is one that quantifies an isolated effect. The questions of interest that relate to targeted effect sizes are narrowly focused and specific, whereas questions of interest that relate to omnibus effect sizes are general. Omnibus effects are generally based on a collection of targeted effects. Thus, whereas omnibus effects describe general effects, targeted effects describe the isolated components that compose the omnibus effect.

To better understand the distinction, consider a balanced single-factor between subjects analysis of variance in which the independent variable is a randomly assigned level of a quantitative factor, such as time spent on a certain task. A trend analysis can be used to decompose the between sums of squares into a set of polynomial trends (e.g., linear, quadratic,

cubic, etc.; see Maxwell & Delaney, 2004, chapter 6, for a review of trend analysis). An effect size (e.g., partial η^2) can be formed for the overall effect (i.e., based on the between sums of squares) or for each of the individual polynomial trends. The partial η^2 value speaks to the overall (i.e., omnibus) effect, whereas the individual partial η^2 values speak specifically to the individual trends (i.e., a targeted effect for each polynomial trend).

Examples of other targeted effect sizes are regression coefficients, pairwise or complex contrasts/comparisons in an analysis of variance context, and path coefficients in a structural equation modeling context. Examples of other omnibus effect sizes are squared multiple correlation, Cramer's V in a chi-square goodness of fit context, and the Mahalanobis distance for a standardized measure of the separation of multiple means (i.e., a vector) for two groups, and fit indices in the context of structural equation modeling.

Some effect sizes are a combination of a targeted and omnibus effect, which we term *semi-targeted* or *semi-omnibus*, both of which have the same meaning. Consider again the trend analysis in the analysis of variance context. Rather than looking at the overall effect or the individual polynomial trends, multiple polynomial trends can be combined (e.g., the cubic and quartic trends) and a partial η^2 can be formed for the multiple trends, creating a semi-targeted effect size. Although for semi-targeted effect sizes the exact contribution of each component is not known, there are fewer unknowns than the corresponding fully omnibus effect size. As another example, consider the increase in the squared multiple correlation coefficient when X_3 and X_4 are added to a model where Y is already modeled as a function of regressors X_1 and X_2 . The contribution of any individual regressor cannot be discerned simply from the change in the squared multiple correlation coefficient when multiple regressors are considered simultaneously. That is, only the aggregate change is known and it is impossible to pinpoint the targeted effect of

each individual regressor without a more targeted effect size. In particular, the change in the squared multiple correlation coefficient for each of the individual regressors could be given, which then provides a targeted effect size.

8. *Effect sizes can be used to convey substantive significance (e.g., clinical, practical, or managerial significance) or for simple description.*

Wedding effect size to substantive significance (e.g., practical, clinical, medical, or managerial importance) may be tempting, but the level of importance attached to a particular value of effect size may vary greatly from one area to another. What is considered impressive in some research areas may not be considered impressive in other research areas. Glass, McGaw, and Smith (1981) argue that “*there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as ‘small,’ ‘moderate,’ ‘large,’ and the like*” (p. 104). Consequently, as tempting as it may be, the idea of linking universal descriptive terms (e.g., “small,” “moderate” or “large”) to specific effect sizes is largely unnecessary and at times misleading (e.g., Baguley, 2009; Lenth, 2001; Robinson et al., 2003; Thompson, 2002).

Our view is that the meaningfulness of an effect is inextricably tied to the particular area, research design, population of interest, and research goal, and it would be inappropriate to wed effect size to some necessarily arbitrary suggestion of substantive significance. This issue is similar to artificially dividing a continuous variable into arbitrary, discrete categories—it is almost always inappropriate to do it (see MacCallum, Zhang, Preacher, & Rucker, 2002 for a review). Our experience is that these qualitative categorizations are sometimes appreciated in certain situations due to their supposed ease of interpretation and even requested or required by reviewers or editors when describing effect sizes. However, categorizing a continuous measure for the purpose of “simpler” interpretation, especially because such categorizations are generally

arbitrary and context specific, is usually a poor way to effectively communicate the results of a study. Furthermore, without clearly communicating the actual value (i. e., not an arbitrary suggestion of meaningfulness) of the effect size, using the study to help plan the sample size for a future study or including such a categorical description of an effect size in a meta-analysis is difficult due to the limited information contained within such a categorical summary.

For example, reducing unplanned absences by 3 days over a year can have a major impact in terms of revenue generation (thus, effect size is -3 here). For example, suppose an employee generates on a mean revenue for an organization of \$70 per hour. For 7 actual working hours for each of three days, there is an increase in overall revenue of \$1,470 ($3 \times 7 \times 70 = 1,470$). But, that is only for a single employee. For multiple employees this revenue can add a significant amount to the company's total revenue. However, increasing a 3rd grader's vocabulary by 3 over the typical vocabulary growth (i.e., the same magnitude in the opposite direction of the revenue generation example) would be rather trivial, since the growth of vocabulary in 3rd grade is generally many words (with 3 words being only a small fraction of the overall vocabulary growth for a typical 3rd grader). These examples can be thought of in the context of Abelson's (1985) paradox, where many small effects can cumulatively have a large effect.

Although effect size can be used to convey substantive significance, it does not necessarily convey something that is important. That is to say, although an effect size (by the definition we provided) is a quantitative reflection of the magnitude of some phenomenon for the purpose of addressing a question of interest, that question may be more descriptive in nature than illustrating something that is necessarily important. For example, it may be of interest to note the mean difference in educational attainment between high-level and mid-level managers in a particular organization. However, that difference may not represent something of substantive importance.

Conveying the substantive importance of something, however, is often a useful purpose of effect size.

As is well known in the literature, statistical significance need not say anything about importance. However, statistical significance has a precise meaning, namely the probability of the observed or more extreme data, given that the null hypothesis is true, is less than the Type I error rate (i.e., α). However, substantive significance does not have such a formal, well-defined meaning, as substantive significance is necessarily context specific and generally a subjective judgment within an area. Cohen reminds us of the distinction between substantive significance and statistical significance when he discusses how researchers sometimes inappropriately interpret “statistical significance” as if the effect size is large or important (1990, p. 1307). What may be clearly a substantively significant effect size in one area may not be substantively significant in another area.

9. Effect sizes values may not be static.

An effect size need not be the current value of an effect; rather, it could be the size of an effect size that once existed (e.g., obtained from historical data) or that may exist in the future (e.g., speculative). Effect sizes can be dynamic and themselves be modeled over time. Whereas a conceptualization of effect size as a static quantity may make sense for some phenomena, many phenomena will change as a function of time, context, or characteristics of the population (e.g., the reliability of a multi-item scale). Additionally, the same effect size value may differ for the same phenomenon across different populations. Correspondingly, careful consideration should be given to the interpretation of effect sizes in the sense that an effect size value in one instance (e.g., time or population) need not be the value obtained in another instance, even if considering the population value, which would not include sampling error.

10. Many combinations of corollaries 1-9 exist.

Not all relevant corollaries of our definition were provided in Corollaries 1-9. Additionally, many of the various corollaries can be combined in multiple ways. Rather than attempt to provide a discussion of combinations of the various corollaries, we focused on what we regard as primary corollaries that may be considered the building blocks for the variety of combinations that exist. For example, an effect size can be a sample value (Corollary 1) that quantifies absolutely (Corollary 2) a dimensionless number (Corollary 3) that is base-rate independent (Corollary 5), observational (Corollary 6), and omnibus (Corollary 7)—such an effect size might be the squared multiple correlation coefficient (R^2) in a multiple regression context. Of course, many other such examples can be given, but the major take-away message is that effect size is a general idea that consists of multiple facets with a multitude of uses.

What Makes a Good Effect Size?

Depending on the situation and the question of interest, some effect size dimensions and effect size measures are preferable to other competing effect sizes. We see as the overarching recommendation that effect sizes (as a general concept encompassing effect size dimension, effect size measure, and effect size value) be tied to the particular research question of interest, which mandates that the question of interest itself be clearly articulated, and the scope and context of the question clearly delineated. Preacher and Kelley (2011) provide a discussion of several desirable properties of effect sizes (points 1-4 below). In particular, they discuss how good effect sizes should have the following properties:

- (1) effect size values should be scaled appropriately, given the measurement and the question of interest;
- (2) effect size values should be accompanied with confidence intervals;

(3) the point estimate of the population effect size value should be independent of sample size;

(4) estimates of effect sizes values should have desirable estimation properties; namely, they should be unbiased (their expected values should equal the corresponding population values), consistent (they should converge to the corresponding population value as sample size increases), and efficient (they should have minimal variance among competing measures).

Not all effect sizes have the complete set of desirable properties. However, competing effect sizes that address the question of interest that most closely satisfy the properties are generally the most preferred. Sometimes historical precedent dictates which effect size or effect sizes is most commonly reported and interpreted. However, nothing bars a researcher from reporting multiple effect sizes addressing the same question of interest in an effort to better communicate the meaning of the results.

Although not necessarily a good or bad quality of an effect size, an appropriate interpretation of an effect size is based on the particular design of the research. Olejnik and Algina (2000, 2003; see also Morris & DeShon, 1997) discuss how an effect size measure in the context of, for example, analysis of variance in a between subjects design can have a different interpretation if a blocking factor is used. When blocking or including covariates, some of what would have been error variance without the blocking factors or covariates may now be explained by the blocking factors or covariates and thus the error variance will be smaller than it otherwise would have been. For a standardized effect size that uses the error variance or some function thereof, it is important to note exactly how the error variance is used. The issue is especially important in the context of meta-analysis when the same question is addressed in multiple studies, yet where some of the studies may use a blocking factor (or covariate). Glass, McGaw, and Smith (1981, chapter 5) suggested that meta-analysts should ignore blocking factors when computing effect

size and provide conversion formulas to convert an effect size based on a blocking factor into one that corresponds to the group difference at the end of a study. Glass et al. (p. 114) argue that an effect size for an unadjusted final status score is more relevant and more readily interpretable.

We emphasize the reporting of confidence intervals for effect sizes throughout this manuscript, and elsewhere. The idea of reporting a confidence interval along with an estimated effect size value is not new (see also Smithson, 2001, 2003, for reviews). However, historically confidence intervals were not often included in many empirical works in psychology and related disciplines. In part due to easy-to-use software and mandates from various sources, confidence intervals are now being used more frequently (e.g., Cumming et al., 2007). The idea of including a confidence interval along with an estimate can be extended to other interval estimates, such as credible intervals (e.g., an analog of confidence intervals from a Bayesian perspective, also termed a posterior interval; e.g., Gelman, Carlin, Stern, & Rubin, 2009), prediction intervals, or tolerance intervals (e.g., Hahn & Meeker, 1991). In some cases the parametric assumptions of confidence intervals will not be satisfied. Even in those situations, confidence intervals with good properties usually can be obtained via a bootstrap method. Bootstrap methods are nonparametric procedures, in which the data obtained are resampled with replacement many times (e.g., 10,000) and the statistic of interest is computed for each of the bootstrap resamples. The statistic of interest from the many bootstrap samples forms an empirical sampling distribution, without reliance on any theoretical distribution (e.g., normal, t , or F distributions). The percentiles (e.g., 2.75 and 97.5 for 95% confidence intervals) or functions of them based on the sampling distribution of the bootstrap resamples can be used to form confidence bounds, even when no known analytic confidence bounds are known (see, for example, Chernick, 2008, or Efron & Tibshirani, 1993, for details).

Any effect size estimated from a sample is itself only an estimate of a corresponding population quantity. Levin (1998) calls the practice of reporting effect sizes with no recourse to hypothesis tests “absurdly pseudoscientific” (p. 45). We agree that reporting *only* an effect size is a scientifically impoverished approach to communicating results; indeed, it would be a throwback to the days before the effects of chance (i.e., sampling error) were considered in understanding obtained results. That is, an effect size alone ignores the sampling distribution of the effect size and thus does not establish a range of plausible parameter values.³ Therefore, a good effect size should have a way to obtain interval estimates, such as a confidence interval.⁴ Interval estimates are critical when a population effect size value is of interest, as such intervals explicitly acknowledge the fallibility of the estimate as a representation of its corresponding population value. In most situations population effect size values are of primary interest, not literally the idiosyncratic effect size value obtained in a particular finite sample (e.g., Balluerka, Gómez, & Hidalgo, 2005; Bird, 2002; Cumming & Finch, 2001; Fidler & Thompson, 2001; Henson, 2006; Kelley, 2005, 2007, 2008; Kelley & Maxwell, 2003; Kirk, 1996; Smithson, 2001; Thompson, 2002, 2007). Consequently, the results of an analysis should always include an interval estimate whenever an effect size estimate is reported, which we regard as a rule that has only one exception we are able to identify, which is when perfectly measured census data (i.e., the whole population) are used.⁵ By including an interval estimate, the interval limits become an explicit part of the interpretation of the results, rather than the point estimate itself, which is known to have error in almost all situations. Thus, seemingly impressive effect size estimates that are accompanied by a wide confidence interval may not seem as impressive if the confidence interval also brackets very unimpressive values.⁶

An important issue when forming an interval estimate for some parameter of interest is whether or not the interval estimate procedure is exact. An exact procedure is one for which the nominal interval coverage (e.g., 95%) is exactly equal to the empirical interval coverage. When certain assumptions are satisfied, many effect sizes have a corresponding exact confidence interval procedure. For those effect sizes that do not yet have an associated exact confidence interval procedure or the assumptions of the confidence interval procedure are not likely to be satisfied sufficiently, researchers often can use various bootstrap techniques to obtain an approximate confidence interval with, in many cases, desirable properties (e.g., Chernick, 2008; Efron & Tibshirani, 1993).

Although confidence intervals are exceedingly important, it is advantageous to have the best point estimate possible. Generally an unbiased estimate of the population quantity is desirable. The bias of an estimator is the degree to which the expected value of the statistic differs from the corresponding population parameter value. Thus, an estimator is unbiased when the expected value of the estimate is exactly equal to the population quantity that it estimates. In general, unbiased estimates are preferred to biased ones, unless a biased statistic is superior in other respects that offset the bias. A biased (or more biased) estimate often can be tolerated in return for substantial gains in precision, such that the overall accuracy of the estimate (the root mean square error) is improved. Statistically, accuracy is a function of precision and bias. If accuracy can be improved (i.e., the root mean square error decreases) by using an estimator that is more biased yet more precise as compared to an unbiased and less precise estimator, the biased estimator can be considered advantageous. This criterion was espoused by Gauss (1809) when he proposed that the smallness of the mean square error of an estimator be used as its measure of excellence (as cited in Thompson, 1968, p. 113). For example, shrinkage estimators, such as

empirical Bayes estimates used in the context of multilevel modeling for estimating individual effects (e.g., an individual's intercept and slope; Kreft & De Leeuw, 1998), are biased estimators that are more precise and generally more accurate than their “unshrunk” counterparts (e.g., Efron & Morris, 1975; Lehmann & Casella, 1998). Additionally, effect sizes that are consistent (i.e., the effect size estimate should converge on the population value as N increases), and efficient (i.e., the effect size estimator should have low sampling variability compared to other estimators of the same quantity) are preferred, all else being equal.

Regarding the desire for an effect size to be independent of sample size, the size of the sample should not affect the expected value of the effect. Rather, an effect size should be an estimate of a particular parameter value. If an effect size is not independent of sample size, then understanding its meaning will be at best difficult. For example, the χ^2 fit statistic is highly dependent on sample size. Some fit indices that use it, such as χ^2/df , are themselves highly dependent on sample size, whereas other fit indices that use the χ^2 are less sensitive to sample size, such as the RMSEA, where the sample size is essentially canceled due to division. Some effect sizes are only mildly sensitive to sample size, in that there is a bias for small sample sizes that disappears for practical purposes when sample size is not small (e.g., Hedges & Olkin, 1985). Such effect sizes (e.g., the standardized mean difference, the squared multiple correlation coefficient, coefficient of variation) are much less problematic than effect sizes whose expectation is highly dependent on sample size. However, it is useful to keep in mind that some effect sizes that are biased due to their dependence on sample size have unbiased (or less biased) versions that can be used (e.g., the adjusted squared multiple correlation coefficient is preferred when estimating the population value as compared to its unadjusted counterpart). Furthermore, an effect size that depends on sample size limits the comparability of such effect sizes across

studies with different sample sizes. The lack of comparability limits the feasibility of a particular effect size from contributing to a cumulative literature in a meta-analytic fashion.

Another consideration when reporting an effect size and its confidence interval is to determine if an unstandardized or a standardized effect size would be a more beneficial way to communicate results. In many cases, it is not difficult to report both the unstandardized and corresponding standardized effect size (see Baguley, 2009, for a different view of reporting both standardized and unstandardized effect sizes). For example, change in the Dow Jones Industrial Average typically is reported in raw value (raw effect size) and as a percentage (a standardized effect size). A practical example is, rather than reporting a correlation matrix with the upper triangle empty (as is often done because the upper triangle is equal to the transposition of the lower triangle), researchers could use the upper triangle to report the covariances with the main diagonal elements equal to the variances of the variables. In such a covariance-correlation table, both standardized and unstandardized values are reported and little additional journal space is required. For another example, regression tables are often reported in journal articles. An additional column for standardized regression coefficients can easily be included. In fact, this is the way in which SPSS reports the output for linear regression analysis, whereas with SAS the option STB in PROC REG yields the same result. Similarly, in a multiple group context, it is generally trivial to report all of the means for several groups, the standard deviation of each group, and the common within group standard deviation (i.e., the root mean square error), as well as selected unstandardized or standardized mean differences and unstandardized or standardized contrasts. When multiple effect sizes (e.g., standardized and unstandardized) are reported, the burden is on the researcher to explain the meaning of the effect sizes. Even if only one type of effect size (standardized or unstandardized) is provided, the other type can generally be

computed if the sufficient statistics have been included as part of the reported results. Sufficient statistics in the context of a particular model are the statistics necessary in order to yield the same results as the complete set of information (i.e., data). For example, in the context of covariance structure analysis, assuming a multivariate normal distribution in the population and known sample size, only the covariance matrix is necessary in order to yield identical results as the full set of data. Including the sufficient statistics can be important so that the work can be included in a meta-analysis at some point or assisting in designing another study in which various effect sizes may be needed.

Discussion

Huberty (2003) and Kirk (1996) remind us that the discussion of effect size measures is not new, but rather has a long history. For example, Yates (1951) noted that researchers paid too much attention to the results of NHSTs and too little attention to the magnitudes of the effects in which they are interested (p. 32; see also, Tukey, 1969). Cohen wrote on the importance of effect sizes: “I have learned and taught that the primary product of a research inquiry is one or more measures of effect size, not p values” (1990, p. 1310; see also Cohen, 1965). Methodologists have long emphasized consideration of effect size when interpreting results and have often stated that researchers have ignored their calls for using effect sizes. In addition to the effect size itself, providing the confidence interval is necessary, because not providing a confidence interval can be considered a disservice to readers (e.g., Bonett, 2008; Kelley, 2005; Thompson, 2002).

We believe the lack of interval estimates, especially confidence intervals, and the lack of interpretation of the interval limits when reported, is a major weakness in research in psychology and related disciplines. Translating the effect size along with the corresponding interval estimate into meaningful substantive terms is something that we see as a principal use of effect sizes.

Some studies report effect sizes but interpret the results from only the perspective of a dichotomous reject or fail-to-reject outcome from a null hypothesis testing framework, perhaps with only an additional consideration of the direction of the effect size.

Our view is that transitioning to a research literature focused on interval estimates of effect sizes that address the question of interest should be a top priority. Confidence intervals are, after all, interval-valued quantifiers of uncertainty. In general, no point estimate in an applied research setting is perfect. Forming an interval estimate for the population effect size is a very useful way of quantifying this uncertainty while simultaneously conveying the plausible range for the parameter of interest at some level of probabilistic uncertainty (e.g., a 95% confidence interval). An implication of such an approach is that hypothesis testing can be conducted by comparing the confidence limits to one or more values of the null hypothesis, if in fact a NHST addresses the question of interest. The only thing that is lost by moving to an interval-centric approach to statistical inference is the exact p -value of the NHST. However, there is no reason why p -values cannot be reported along with effect sizes and confidence intervals. We believe that even when the null hypothesis is not rejected, estimates of effect size and confidence intervals are valuable and should be reported (contrast this view with Knapp & Sawilowsky, 2001a, 2001b). Indeed, effect sizes, regardless of their associated p -values, can be used as input for future meta-analysis. In some situations effect sizes are biased, especially in small samples. Often such bias can be reduced by using an unbiased (or more unbiased) estimate, which is discussed in some meta-analytic sources (e.g., Hedges & Olkin, 1985). Systematically avoiding the publication of effect sizes for effects that failed to reach statistical significance can lead to publication bias, where truth and published reality differ (e.g., Rostein, Sutton, & Borenstein, 2005).

Effect sizes and research design go hand-in-hand, especially issues of sample size planning. Methods for sample size planning can be broadly segmented into answering two fundamentally different types of research questions about effect sizes: the first involves inferring the existence of a non-null effect in the population (power analysis) and the second involves inferring the magnitude of the size of the effect (accuracy in parameter estimation; e.g., see Maxwell, Kelley, & Rausch, 2008, for a recent review of the two approaches to sample size planning). In particular, when the research question of interest concerns rejecting a null hypothesis, it is advisable to plan a study so that there is sufficient statistical power to reject the null hypothesis (e.g., Bausell & Li, 2002; Chow, Shao, & Wang, 2003; Cohen, 1988; Kraemer & Thiemann, 1987). However, when the research question of interest concerns the magnitude of an effect in the population, it is advisable to plan a study so that the obtained confidence interval will be sufficiently narrow, which is the goal of the accuracy in parameter estimation approach to sample size planning (e.g., Jiroutek, Muller, Kupper, & Stewart, 2003; Kelley & Maxwell, 2003). Regardless of the approach, sample size planning procedures generally require effect size values in order to implement the procedure, which are usually either speculated (i.e., population value assumed known) or set to some minimum value of practical importance. Due to the difficulty when estimating population effect size value(s) for sample size planning purposes, effect sizes have been called the “problematic parameter” (Lipsey, 1990, chapter 3). Better reporting of effect size values from individual studies will facilitate future sample size planning in a similar way that meta-analysis benefits.

The use of the term effect size separates what is simply a statistic (or parameter) into a statistic (or parameter) that addresses a question of interest for some purpose. In some ways, however, effect size is simply a name applied to special types of statistics and parameters, but,

“what’s in a name?” In Shakespeare’s *Romeo and Juliet*, Juliet notes “that which we call a rose by any other name would smell as sweet.” Correspondingly, whatever the field calls the idea of a quantitative reflection of the magnitude of some phenomenon, in some ways, is irrelevant. It is when we use this quantitative measure to address a question of interest that advances can be made. Using quantitative measures in such a way is a rationale for the widespread use of effect sizes, which we hope to help advance by having a wide ranging discussion on what the term effect size conveys and how it encompasses a broad set of statistics (or parameters) when they are used to address questions of interest.

We believe that with a full implementation of the effect size movement in the applied literature, study results will be better communicated and studies can be better planned in an effort to increase the cumulative knowledge of psychology and related disciplines. A fundamental question is “How do we learn from data?” Certainly calling something an effect size does not necessarily help us learn from data. However, understanding how different statistics can be used to estimate the magnitude of a phenomenon, where the magnitude helps to address a question of interest, we would argue, is an important way to learn from data. We believe that our encompassing definition and delineation of effect size will help advance applied research by elucidating what is actually meant by a term often mentioned but not often clearly articulated.

References

- Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129-133.
- Aiken, L. R. (1994). Some observations and recommendations concerning research methodology in the behavioral sciences. *Educational and Psychological Measurement*, *54*, 848-860.
- Altman, D. G. (1998). Confidence intervals for the number needed to treat. *British Medical Journal*, *317*, 1309-1312.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC.
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC.
- Association for Psychological Science. (2011). Submission guidelines for *Psychological Science*, Accessed http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions on 3/7/2011.
- Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603-617.
- Balluerka, N. Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology*, *1*, 55-70.
- Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences*. New York, NY: Cambridge.
- Berry, K. J., & Mielke, P. W., Jr. (2002). Least sum of Euclidean regression residuals: Estimation of effect size. *Psychological Reports*, *91*, 955-962.

- Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62*, 197-226.
- Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13*, 99-109.
- Bridgman, P. W. (1922). *Dimensional analysis*. New Haven, CT: Yale University Press.
- Carman, R. A. (1969). *Numbers and units for physics*. John Wiley & Sons: New York, NY.
- Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.
- Chow, S.-C., Shao, J., & Wang, H. (2003). *Sample size calculations in clinical research*. New York, NY: Taylor & Francis.
- Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.). *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.
- Cohen, J., Cohen, P., West, S. G., Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.
- Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230-232.
- Cumming, G., & Finch, (2001). A primer on the understanding, use, and calculation of

- confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 530-572.
- Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, 94, 361-366.
- Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalization. *Journal of the American Statistical Association*, 70, 311-319.
- Efron, B., & Tibshirani, R. W. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.
- Ellis, B. (1966). *Basic concepts of measurement*. New York, NY: Cambridge.
- Ellis, P. D. (2010). *The essential guide to effect sizes statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15, 119-126.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect Sizes. *Educational and Psychological Measurement*, 61, 575-604.
- Ford, G. R., & Cullmann, R. E. (1959). *Dimensions, units, and numbers in the teaching of physical sciences*. New York, NY: Teachers College, Columbia University.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2009). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

- Grissom, R. J., & Kim, J. J. (2005). *Effect size for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
- Grissom, R. J., & Kim, J. J. (2012). *Effect size for research: Univariate and multivariate applications* (2nd ed). New York, NY: Routledge.
- Hahn, G. J., & Meeker, W. Q. (1991), *Statistical intervals: A guide for practitioners*, New York: NY: Wiley.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.
- Henson, R. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, 34, 601-629.
- Huberty, C. (2003). Multiple correlation versus multiple regression. *Educational and Psychological Measurement*, 63, 271-278.
- International Committee of Medical Journal Editors (2007). *International Committee of Medical Journal Editors' uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication*. Available online: http://www.icmje.org/2007_urm.pdf, accessed August 15, 2010.
- Ipsen, D. C. (1960). *Units, dimensions, and dimensionless numbers*. New York, NY: McGraw-Hill.
- Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, 59, 580-90.
- Kazis, L. E., Anderson, J. J., and Meenan, R. F. (1989). Effect Sizes for Interpreting Changes in Health Status. *Medical Care*, 27, S178-S189.
- Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the

- standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65, 51-69.
- Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20(8), 1-24.
- Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research*, 43, 524-555.
- Kelley, K., & Maxwell, S. E. (2003). Sample size planning for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8, 305-321.
- Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, 65, 3-5.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Kirk, R. E. (2002). Experimental design. In I. B. Weiner (Series Ed.) and J. Schinka & W. F. Velicer (Vol. Eds.) *Handbook of psychology* (pp. 3-32). New York, NY: Wiley.
- Knapp, T. R., & Sawilowsky, S. S. (2001a). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, 70, 65-79.
- Knapp, T. R., & Sawilowsky, S. S. (2001b). Strong arguments: Rejoinder to Thompson. *The Journal of Experimental Education*, 70, 94-95.
- Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?: Statistical power analysis in research*. Newbury Park, CA: Sage.
- Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA; Sage.
- Langhaar, H. L. (1951). *Dimensional analysis and theory of models*. New York, NY: Chapman

& Hall.

Lehmann, E. L., Casella, G. (1998). *Theory of point estimation* (2nd edition). New York, NY:

Springer.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The*

American Statistician, 55, 187-193.

Levin, J. R. (1998). What if there were no more bickering about statistical significance tests?

Research in the Schools, 5, 43-53.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury

Park, CA: Sage.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of

dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model*

comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power

and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537-563.

McGrath, R. E., Meyer, G. J. (2006). When effect sizes disagree: The case of r and d

Psychological Methods, 11, 386-401.

Miller, S. (2007). *Developmental research methods* (3rd ed.), Thousand Oaks, CA: Sage.

Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis

of variance for use in meta-analysis. *Psychological Methods*, 2, 192-199.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, 82, 3-5.

Nakagawa, S., & Cuthill, I. (2007). Effect size, confidence interval and statistical significance: a

practical guide for biologists. *Biological Reviews*, 82, 591-605.

National Center for Education Statistics. (2002). *NCES statistical standards* (revised ed.).

Washington, DC: Department of Education.

- Olejnik, S., & Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. *Contemporary Educational Psychology* 25, 241–286.
- Olejnik, S., & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8, 434-447.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed). Cambridge University Press: New York, NY.
- Peyton, V. (2005). Effect size. In S. W. Lee (Ed.), *Encyclopedia of school psychology* (pp. 186-187). Thousand Oaks, CA: Sage.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93-115.
- Robinson, D., Whittaker, T., Williams, N., & Beretvas, S. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *Journal of Experimental Education*, 72, 51-64.
- Rosenthal, R. (1994) Parametric measures of effect size. In H. Cooper and L. V. Hedges (Eds.) *The handbook of research synthesis* (pp. 231-244). New York, NY: Russell Sage Foundation.
- Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166-169.
- Rostein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Hoboken, NY: Wiley.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1, 115-129.
- Schulz, K. F., Altman, D. G., Moher, D., for the CONSORT Group (2010). CONSORT 2010

- Statement: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal*, 340, 697-702.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, 61, 605-632.
- Smithson, M. (2003). *Confidence intervals*. Sage: Thousand Oaks, CA.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Society for Industrial Organizational Psychology.
- Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine*, 318, 262-264.
- Task Force on Reporting of Research Methods in AERA Publications (2006). *Standards for reporting on empirical social science research in AERA publications*. Washington, DC: American Educational Research Association.
- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31, 25-32.
- Thompson, B. (2004). The "significance" crisis in psychology and education. *Journal of Socio-Economics*, 54, 607-613.
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423-432.

- Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, *63*, 113-122.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83-91.
- Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*, 473-481.
- Vaske, J. J., Gliner, J. A., & Morgan, G. A. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife*, *7*, 287-300.
- Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.
- Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 19-34.

Footnotes

¹ The previous (5th) edition of the *APA Manual* (2001) did state that confidence intervals “can be an extremely effective way of reporting results” (p. 22) and that researchers should “provide the reader not only with information about statistical significance, but also with enough information to assess the magnitude of the observed effect or relationship” (p. 25).

² There is an element of arbitrariness in selecting the Type I error rate. Although a Type I error rate of .05 may be the modal value used in psychology and related disciplines, the value of .05 is itself simply an arbitrary convention.

³ Another implication of ignoring the sampling distribution of the effect size is that it is unclear whether the population effect size differs from some specified null value. That is, by ignoring the sampling distribution no NHSTs can be performed. Although we generally advocate the use of confidence intervals, under some circumstances NHSTs can be beneficial.

⁴ If inference is done in the context of a Bayesian framework, credible intervals based on both the specified priors and data should be provided.

⁵ Even for census data an interval estimate should often be provided, as populations are often dynamic and interpretations can then apply more broadly than being wedded to the particular population for which data were collected at a particular point in time.

⁶ We have purposely used the term “impressive” here rather than “large” effect size. The reason is because an effect size may be considered “small” by some standard but in reality may be very “impressive.” Indeed, in some cases effect sizes are most impressive by being as small as possible.