Psychology in the Schools, Vol. 44(5), 2007
Published online in Wiley InterScience (www.interscience.wiley.com).

© 2007 Wiley Periodicals, Inc.
DOI: 10.1002/pits.20234

EFFECT SIZES, CONFIDENCE INTERVALS, AND CONFIDENCE INTERVALS FOR EFFECT SIZES

BRUCE THOMPSON

Texas A&M University, Baylor College of Medicine

The present article provides a primer on (a) effect sizes, (b) confidence intervals, and (c) confidence intervals for effect sizes. Additionally, various admonitions for reformed statistical practice are presented. For example, a very important implication of the realization that there are dozens of effect size statistics is that *authors must explicitly tell readers what effect sizes they are reporting*. With respect to confidence intervals, when interpreting a 95% interval, we should *never say that we are 95% confident that our interval captures the estimated population parameter*. It is explained that *effect sizes should be reported even for statistically nonsignificant effects*. And, most importantly of all, it is emphasized that effect sizes should *not* be interpreted using Cohen's benchmarks. Instead, we ought to *interpret our effects in direct and explicit comparison against the effects in the related prior literature*. © 2007 Wiley Periodicals, Inc.

The origins of null hypothesis statistical significance tests (NHSST) can be traced back to the 1700s, but in the early 1900s a wide array of NHHST tests (e.g., Student's *t*, ANOVA) first became available to the emerging field of psychology (Huberty & Pike, 1999). Yet, the widespread uptake by psychologists of NHSST did not occur until the 1950s (Hubbard & Ryan, 2000). And almost from the beginning, overreliance on NHHST attracted critics, such as Boring (1919) and Berkson (1938).

During the last 15 years, the frequency of published criticisms of NHSST has grown exponentially and across diverse disciplines, as reported in one of the graphs presented by Anderson, Burnham, and Thompson (2000) in their *Journal of Wildlife Management* article. These criticisms have been published, for example, in the fields of biology (Suter, 1996; Yoccuz, 1991), economics (Thompson, 2004; Ziliak & McCloskey, 2004), education (Carver, 1978; Thompson, 1996), psychology (Cohen, 1994; Schmidt, 1996), and the wildlife sciences (Johnson, 1999).

The tenor, if not the substance, of these criticisms can be conveyed via some quotations. For example, Schmidt and Hunter (1997) argued that, "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (p. 37). Rozeboom (1997) was equally direct, stating: "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students. . . . [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism" (p. 335).

In consideration of such views, the American Psychological Association (APA) Board of Scientific Affairs in 1996 appointed a Task Force on Statistical Inference (Task Force) to recommend whether or not NHSST should be banned from APA journals. Ultimately, the Task Force did not recommend an NHSST ban, but did encourage a variety of statistical reforms, such as reporting effect sizes and confidence intervals (Wilkinson & Task Force on Statistical Inference, 1999).

The purpose of the present article is to provide a primer on (a) effect sizes, (b) confidence intervals, and (c) confidence intervals for effect sizes. For book-length treatments, the reader is directed to Grissom and Kim (2005), Kline (2004), and Thompson (2006a).

Effect Sizes

Effect size statistics (e.g., Cohen's *d*, Glass' Δ, η^2, adjusted r^2 or adjusted R^2, ω^2) characterize the extent to which sample results diverge from the expectations specified in the null hypothesis (Cohen, 1994; Vacha-Haase & Thompson, 2004). For example, for the H_0:

Correspondence to: Bruce Thompson, TAMU Department of Educational Psychology, College Station, TX 77843-4225. E-mail: bruce-thompson@tamu.edu

$mdn_{\text{FRESHMEN}} = mdn_{\text{SOPHOMORES}} = mdn_{\text{JUNIORS}} = mdn_{\text{SENIORS}}$, effect sizes are zero if the sample statistics are 5.0, 5.0, 5.0, and 5.0, nonzero if the sample medians are 5.0, 5.0, 5.0, and 5.1, and greater still if the sample statistics are 5.0, 6.0, 7.0, and 8.0. Effect sizes are not new, as Huberty (2002) noted in his history. However, until recently textbooks gave short shrift to effect size treatments, and thus many researchers were not exposed to these estimates as part of older doctoral curricula (Capraro & Capraro, 2002).

The APA Task Force on Statistical Inference emphasized that effect sizes (e.g., Cohen's d, ω^2, η^2) should "*always*" be reported along with p values, and that "reporting and interpreting effect sizes in the context of previously reported effects is *essential* to good research" (Wilkinson & Task Force on Statistical Inference, 1999, p. 599, emphasis added). In response, the fifth edition of the APA (2001) *Publication Manual of the American Psychological Association* (*Publication Manual*), used by more than 1000 journals, included a new declaration that, "For the reader to fully understand the importance of your findings, it is *almost always necessary* to include some index of effect size or strength of relationship in your results section" (pp. 25–26, emphasis added), and labeled the "failure to report effect sizes" as a "defect in the design and reporting of research" (APA, 2001, p. 5).

However, given the limitations of the *Publication Manual* (Fidler, 2002), and the inherent barriers of these few words rising to prominence (or at least consciousness) given their placement within a 439-page tome, two dozen journals have gone further and now explicitly *require* effect size reporting. Two of these journals have subscriptions greater than 50,000 and are the organizational "flagship" journals of the Council for Exceptional Children and the American Counseling Association. Indeed, Fidler noted that, "Of the major American associations, only all the journals of the American Educational Research Association have remained silent on all these issues" (p. 754). As Grissom and Kim (2005) emphasized, "NHSST does not sufficiently indicate how much better the superior treatment is or how strongly the variables are related. . . . If two treatments are not equally effective, the better of the two can be anywhere from slightly better to very much better than the other" (p. 2). Grissom and Kim (2005) also noted that readers "have a right to see estimates of effect sizes. Some might even argue that not reporting such estimates in an understandable manner . . . may be like withholding evidence" (p. 5).

There are 40+ effect sizes, and the number is growing (Kirk, 1996). More exotic (but promising) effect sizes include Huberty's group overlap index I (Hess, Olejnik, & Huberty, 2001; Huberty & Holmes, 1983; Natesan & Thompson, 2007) and Grissom's (1994) "probability of superiority." In psychology, we tend to use effect sizes that have been standardized (i.e., by removing the *SD* of the measures from the effect size via division), so that we can compare effect sizes involving a construct (e.g., self-concept) apples-to-apples even when researchers use different measures of the construct.

Generally, effect sizes can be converted into each other's metrics using well-known formulas (see Thompson, 2002a). A very important implication of the realization that there are dozens of effect size statistics is that *authors must explicitly tell readers what effect sizes they are reporting*, so that the effects can be properly interpreted and compared apples-to-apples across studies!

Various frameworks can be used for categorizing effect sizes (cf. Kirk, 1996; Thompson, 2006b). Here the most commonly used effect sizes are presented in (a) a standardized difference versus variance-accounted-for and (b) an uncorrected for bias versus corrected framework.

Uncorrected, Standardized Differences

Two of the most commonly used effect sizes are Cohen's d and Glass's Δ, both of which are standardized mean differences. For equal group sizes, Cohen's d can be computed as

$$d = (M_{\text{E}} - M_{\text{C}})/\text{SQRT}[(SD_{\text{E}}^2 + SD_{\text{C}}^2)/2], \tag{1}$$

where M_E and SD_E are the posttest mean and *SD* within the experimental group and M_C and SD_C are the posttest mean and *SD* within the control group. For unequal group sizes, weighted average *SD*s are computed. Similarly, Glass' Δ is computed as

$$\Delta = (M_E - M_C)/SD_C. \tag{2}$$

Glass's Δ is most useful when group sizes are quite large, and concerns exist as to whether the intervention may have affected not only M_E, but also SD_E. However, when concerns about intervention effects on posttest dispersion are less noteworthy, Cohen's *d* has the advantage of greater precision in estimating the denominator of these standardized effects, because total *n* is larger if both groups are used to estimate a pooled *SD*.

Certainly, alternatives to these effects can be conceptualized. For example, a standardized median difference can be computed as

$$d_{\text{mdn}} = (mdn_E - mdn_C)/\text{SQRT}[(SD_E^2 + SD_C^2)/2]. \tag{3}$$

This effect may be useful when posttest data are skewed or researchers want estimates to be less sensitive to outliers.

Uncorrected, Variance-Accounted-For Effects

Because *all* parametric analyses are correlational and part of a single General Linear Model (GLM) (cf. Cohen, 1968; Knapp, 1978; Thompson, 2006a), all parametric analyses (e.g., *t* tests, ANOVA, ANCOVA, MANOVA, descriptive discriminant analysis) yield effect sizes analogous to the Pearson r^2. For example, in ANOVA η^2 can be computed as

$$\eta^2 = SOS_{\text{EXPLAINED}}/SOS_{\text{TOTAL}}. \tag{4}$$

Of course, given the GLM, the same formula can be used to compute the Pearson r^2 or the multiple R^2.

These related estimates are all interpreted in the same manner. For example, an η^2 of 10% means that, given knowledge of group membership on the independent variable, we can explain 10% of the variability of the outcome variable.

Corrected, Variance-Accounted-For Effects

Ordinary Least Squares (OLS) analyses (e.g., *t* tests, ANOVA, ANCOVA, MANOVA, descriptive discriminant analysis) estimate effect sizes (e.g., η^2, r^2, R^2) by fitting an analytic model (e.g., *t* test, ANOVA, regression) to sample data. The estimated effect will generalize well to the population if (and only if) the sample data are representative of the population.

Unfortunately, all samples (like all people) tend to have their own personalities and idiosyncrasies. Thus, all sample-based estimates of population effect sizes (e.g., η^2, r^2, R^2) tend to be *positively biased* (i.e., overestimate true population effects), because the OLS analyses cannot discriminate between variability in the sample data that are real (i.e., represent true population variability) and sample data that are unique to a given sample (i.e., does not exist in the population, or in any other sample, each of which has its own unique idiosyncrasies).

Fortunately, we know that sample data have more idiosyncrasies (i.e., more "sampling error variance") when (a) sample size is small, (b) the number of measured variables is large, and (c) population effect size is small. The reasons why these three factors impact sampling error are beyond the scope of the present brief treatment, but the interested reader is referred to Thompson (2002b, 2006a).

The Ezekiel correction is one formula that can be applied with both multiple R^2 and the Pearson r^2 (Wang & Thompson, 2007). The corrected R^2, R^{2*}, can be computed as

$$R^{2*} = 1 - [(n-1)/(n-p-1)]\,[1-R^2], \tag{5}$$

where p is the number of predictor variables. SPSS will compute this adjusted effect size automatically if we run the REGRESSION analysis command.

For example, for $r = 0.10$, $r^2 = 0.01$, and

$$\begin{gathered} 1 - [(10-1)/(10-1-1)]\,[1-0.01] \\ 1 - [9/8]\,[0.99] \\ 1 - [1.12]\,[0.99] \\ 1 - 1.11 \\ r^{2*} = -0.11. \end{gathered}$$

Obviously, negative corrected variance-accounted-for effect sizes are troubling and indicate that all the detected effects (i.e., $r^2 = 1\%$), and more, are an artifact of sampling error variance.

The formulas can also be used to estimate the expected sample R^2 (or r^2) if the population effect size is zero, given the design. For example, for a regression study with $n = 15$ and $p = 5$, if the population R^2 is 0, the expected sample R^2 is 0.357, because using Formula 5 we have

$$R^{2*} = 1 - [(10-1)/(10-5-1)]\,[1-0.357] = 0.000.$$

Thus, if we want a positive adjusted R^2 when we are about to do a study with five predictor variables and $n = 10$, we had better get a sample R^2 of at least 0.357, or alternatively we had better use a larger sample size.

For ANOVA, we can compute Hays' (1981) ω^2 as

$$\omega^2 = [SOS_{\mathrm{BETWEEN}} - (k-1)MS_{\mathrm{WITHIN}}]/[SOS_{\mathrm{Y}} + MS_{\mathrm{WITHIN}}], \tag{6}$$

where k is the number of levels in the ANOVA way, SOS_{BETWEEN} is the sum of squares between, and MS_{WITHIN} is the mean square within or error. For example, if we have a fixed-effects one-way ANOVA with eight participants in each of four groups, we would obtain η^2 and ω^2 for the following results:

Source	*SOS*	*df*	*MS*	*F*	η^2	ω^2
"A" way	15	3	5.00	4.00	30.00%	21.95%
Error	35	28	1.25			
Total	50	31	1.61			

SPSS will output ω^2 as an optional ANOVA statistic, upon request.

Corrected, Standardized Differences

Because Cohen's d estimates effect size using sample data and OLS estimation, this effect size also tends to be positively biased (Hedges, 1981, 1982). Thus, a corrected d, d^*, can be computed as

$$d^* = [1 - 3/[4(n_1 + n_2 - 2)] - 1]\,d \tag{7}$$

according to Rosenthal's (1994) Equations 16-32 and 16-33.

Confidence Intervals for Statistics

The APA Task Force strongly encouraged the use of confidence intervals, noting that

> Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability [or the lack thereof] across studies. . . . Collecting intervals across studies also helps in constructing plausible regions for population parameters. (Wilkinson & Task Force on Statistical Inference, 1999, p. 599)

And the 2001 *Publication Manual* of the American Psychological Association suggested that confidence intervals (CIs) represent “in general, *the best* reporting strategy. The use of confidence intervals is therefore *strongly recommended*” (p. 22, emphasis added).

However, empirical studies confirm that confidence intervals for statistics are virtually never reported in the social sciences (Finch, Cumming, & Thomason, 2001). For example, Kieffer, Reese, and Thompson (2001) coded roughly 1300 articles published in 10 volumes of the *American Educational Research Journal* and the *Journal of Counseling Psychology* and found that less than a handful of articles reported CIs. And, as Thompson (2002b) suggested, “It is conceivable that some researchers may not fully understand statistical methods that they (a) rarely read in the literature and (b) infrequently use in their own work” (p. 26; also see Cumming & Finch, 2005).

Indeed, users of CIs often misinterpret intervals (e.g., 95% confidence intervals) as reflecting certainty (i.e., 100%) that their particular confidence interval captures the population parameter. Figure 1 presents a series of 95% confidence intervals for the mean for 20 random samples, each $n = 12$, drawn from a population with $\mu = 50$ and $\sigma = 10$. The figure was drawn with the Exploratory Software for Confidence Intervals software (ESCI; see Cumming & Finch, 2001).

Note that the 14th of the 20 intervals, with a mean of approximately 56, did not capture the population parameter of $\mu = 50$. The failure of some 95% intervals to capture the true parameter is expected. Indeed, what computing 95% CIs for a statistic means is that, if we drew infinitely many random samples from the population, exactly 95% of the CIs would capture the parameter, and exactly 5% would not.

In short, the 95% confidence statement does not apply to a single interval. According to Formula 3 in Thompson’s (2006b) chapter, which has come to be called “Thompson’s inequality,” we must always remember that

$$1 \neq \infty.$$

So, when interpreting a 95% interval, we should *never say that we are 95% confident that our interval captures the estimated population parameter*. Nevertheless, CIs are extremely useful, because they convey not only our point estimate, but also, via the width of the intervals, something about the precision of our estimates. And, across a series of CIs from a body of studies, as Schmidt (1996) so astutely noted, we will eventually obtain a reasonable estimate of a population parameter, even if our initial expectations are wildly wrong.

Confidence Intervals for Effect Sizes

Formulas can be used to obtain (a) statistics, (b) CIs for statistics, and (c) effect sizes. However, formulas are less useful in obtaining CIs for effect sizes, and, instead, computer-intensive iteration procedures must be invoked (see Cumming & Finch, 2001). Basically, iteration successively guesstimates CIs for effect sizes until a statistically reasonable CI is estimated. Happily, for many research situations, the required software is available to run stand-alone on a microcomputer (e.g., Steiger & Fouladi, 1992), in Excel (Cumming & Finch, 2001), in SPSS (Smithson, 2001), or in SAS (Algina & Keselman, 2003; Algina, Keselman, & Penfield, 2005b). This is an area of

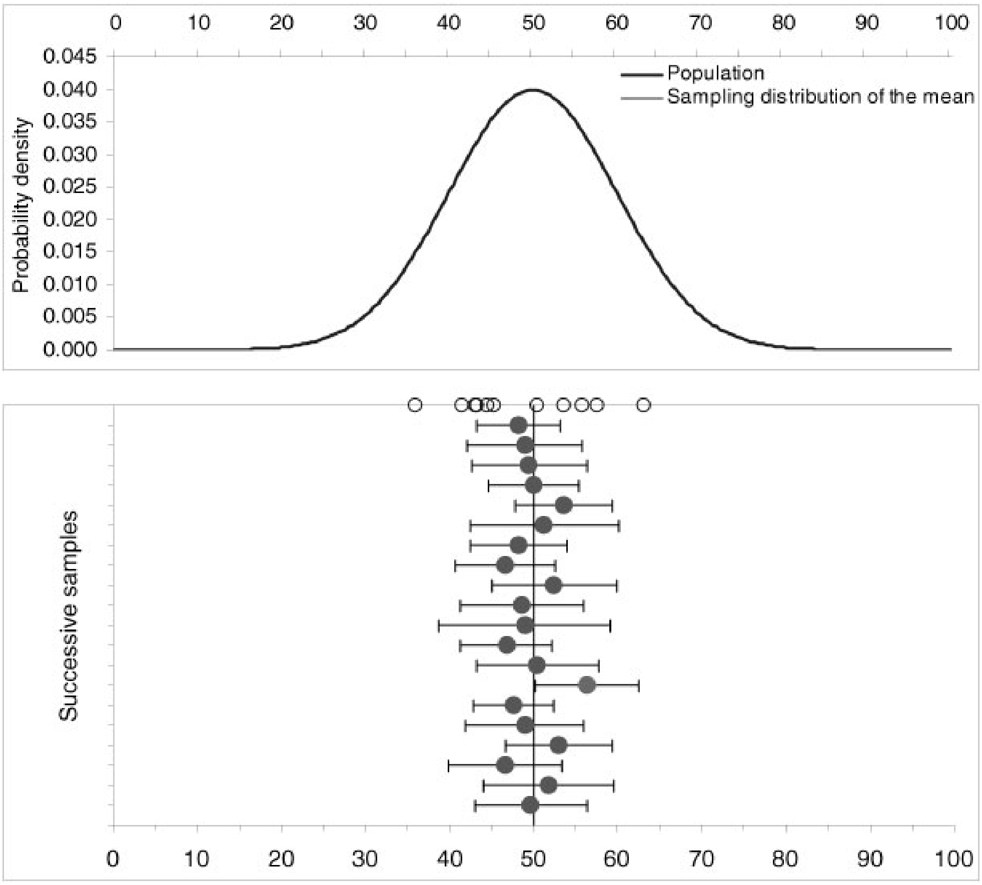

FIGURE 1. Ninety-five percent confidence intervals for the mean for 20 random samples ($n = 12$), from a population with $\mu = 50$ and $\sigma = 10$. The last sample of 12 scores is represented using circles at the top of the "Successive Samples" box, with the mean and the 95% CI for the mean presented immediately below the 12 scores. For the remaining 19 samples, only the means and 95% CIs about the means are presented.

quickly developing theory and methodology (e.g., Algina, Keselman, & Penfield, 2005a; Keselman, Algina, & Fradette, in press).

Table 1 presents a hypothetical literature for studies investigating treatment effects for a new intervention against childhood depression where means of the treatment groups are tested against the null that they equal a parameter diagnostic cutoff score. This hypothetical literature honors Rosnow and Rosenthal's (1989) view that "surely, God loves the .06 [level of statistical significance] nearly as much as the .05" (p. 1277).

Of course, in a literature plagued by the "file drawer" problem (Rosenthal, 1979), only the negative results in the first study ($p < .05$) would be published, and failures to replicate, in the form of the remaining 10 nonstatistically significant but replicable positive outcomes, would be less likely to be published, or even submitted (see Greenwald, 1975). Figure 2 illustrates that, from such a literature as a whole, assuming all studies were somehow published, a different picture might emerge. This portrait suggests generally positive treatment effects that are statistically significant across the literature as a whole.

Table 1
Hypothetical Literature Consisting of 11 Studies

				95% CI for *d*	
Study	*d*	*n*	$p_{\text{CALCULATED}}$	Lower	Upper
1	−0.45	22	0.047	−0.88	0.00
2	0.30	40	0.065	−0.02	0.61
3	0.70	9	0.069	−0.05	1.42
4	0.65	10	0.070	−0.05	1.32
5	0.35	30	0.065	−0.02	0.71
6	0.50	16	0.064	−0.03	1.01
7	0.35	31	0.061	−0.01	0.71
8	0.60	12	0.062	−0.03	1.21
9	0.40	24	0.062	−0.02	0.81
10	0.45	19	0.065	−0.03	0.92
Combined	0.327	213	<.001	0.19	0.46
New study	0.52	15	0.064	−0.03	1.05
New combined	0.339	228	<.001	0.21	0.47

The Figure 2 literature also emphasizes the important implication that *effect sizes should be reported even for statistically nonsignificant effects*. Quite noteworthy pooled effects (e.g., meta-analytic effects) can arise even from a literature where all effects within single studies are not statistically significant. It is also important to remember Schmidt's (1996) wise admonition that, "Meta-analysis . . . has revealed how little information there typically is in any single study. It has

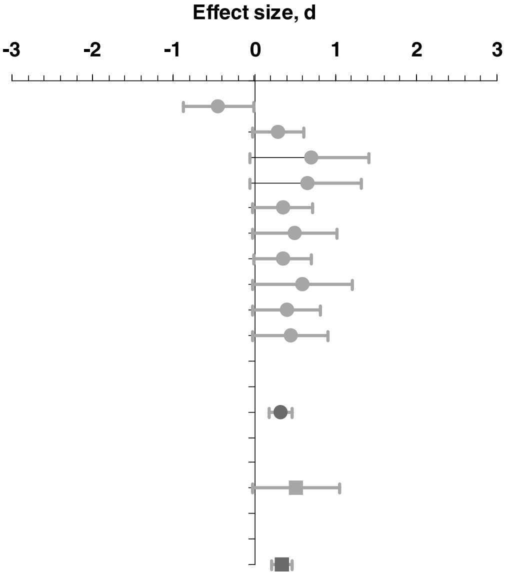

FIGURE 2. Hypothetical literature consisting ultimately of 11 studies.

shown that, contrary to widespread belief, a single primary study can rarely resolve an issue or answer a question" (p. 127).

Discussion

Internationally respected statistician Roger Kirk (2003) painted a portrait of a possible future for psychological science, a world in which effect sizes play a pivotal role:

> It is evident that the current practice of focusing exclusively on a dichotomous reject–nonreject decision strategy of null hypothesis testing can actually *impede scientific progress*. . . . In fact, focusing on *p* values and rejecting null hypotheses actually *distracts us from our real goals*: deciding whether data support our scientific hypotheses and are practically significant. The focus of research should be on our scientific hypotheses, what data tell us about the magnitude of effects, the practical significance of effects, and the steady accumulation of knowledge. (Kirk, 2003, p. 100, italics added)

In short, we need to ask (a) whether our effects are noteworthy from a practical point of view, and (b) to whom our effect size results generalize.

Of course, we must recognize that if we violate the assumptions of statistical methods (e.g., homogeneity of variance in ANOVA or homogeneity of regression in ANCOVA), we compromise not only our *p* values but also our effect estimates. And our effect sizes do not generalize beyond the limits of our research designs (Olejnik & Algina, 2003). If a literature evaluates treatment efficacy over therapy sessions of 45, 50, or 55 min in length and finds that longer is better, our results simply do not speak to the question of whether 90-min sessions would be still better.

And, most importantly of all, we must *not* fall into using Cohen's benchmarks for "small," "medium," and "large" effects, which can be expressed in several effect size metrics. Cohen (1988) himself emphasized that "these proposed conventions were set forth throughout with much diffidence, qualifications, *and invitations* ***not*** *to employ them if possible*. . . . They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of [effect size] magnitude" (p. 532, italics added). As noted elsewhere, "if people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (Thompson, 2001, pp. 82–83).

As Thompson (2002b) emphasized, we ought *to interpret our effects in direct and explicit comparison against the effects in the related prior literature*. This forces us to look at effects in context and to evaluate the precision and replicability of effects within a literature. Large effects in some contexts are quite trivial. Conversely, numerically very small effects, in at least some contexts, can be extremely noteworthy. For example, small effects for interventions against highly robust disorders (e.g., eating disorders) may be important (Prentice & Miller, 1992). And small effects that work incrementally over time can be extremely important. For example, a kindergarten compensatory educational intervention might shift slightly the group mean in the intervention group but also may shift the learning curve itself and thus the rate at which children learn throughout their entire educational careers. Thus, the cumulative impacts of small changes over time may be huge, even when the initial impacts seem small!

References

Algina, J., & Keselman, H.J. (2003). Approximate confidence intervals for effect sizes. Educational and Psychological Measurement, 63, 537–553.

Algina, J., Keselman, H.J., & Penfield, R.D. (2005a). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. Psychological Methods, 10, 317–328.

Algina, J., Keselman, H.J., & Penfield, R.D. (2005b). Effect sizes and their intervals: The two-level repeated measures case. Educational and Psychological Measurement, 65, 241–258.

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

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management, 64, 912–923.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the Chisquare test. Journal of the American Statistical Association, 33, 526–536.

Boring, E.G. (1919). Mathematical vs. scientific significance. Psychological Bulletin, 16, 335–338.

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

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378–399.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426–443.

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

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997–1003.

Cumming, G., & Finch, S. (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, 532–574.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist, 60, 170–180.

Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. Educational and Psychological Measurement, 62, 749–770.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. Educational and Psychological Measurement, 61, 181–210.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1–20.

Grissom, R.J. (1994). Probability of the superior outcome of one treatment over another. Journal of Applied Psychology, 79, 314–316.

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

Hays, W.L. (1981). Statistics (3rd ed.). New York, NY: Holt, Rinehart and Winston.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107–128.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490–499.

Hess, B., Olejnik, S., & Huberty, C.J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. Educational and Psychological Measurement, 61, 909–936.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. Educational and Psychological Measurement, 60, 661–681.

Huberty, C.J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62, 227–240.

Huberty, C.J., & Holmes, S.E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15–26.

Huberty, C.J., & Pike, C.J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 1–23). Stamford, CT: JAI Press.

Johnson, D.H. (1999). The insignificance of statistical significance testing. Journal of Wildlife Management, 63, 763–772.

Keselman, H.J., Algina, J., & Fradette, K. (in press). Robust confidence intervals for effect size in the two-group case. Journal of Modern Applied Statistical Methods.

Kieffer, K.M., Reese, R.J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. Journal of Experimental Education, 69, 280–309.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746–759.

Kirk, R.E. (2003). The importance of effect magnitude. In S.F. Davis (Ed.), Handbook of research methods in experimental psychology (pp. 83–105). Oxford, UK: Blackwell.

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

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance-testing system. Psychological Bulletin, 85, 410–416.

Natesan, P., & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. Educational and Psychological Measurement, 67, 59–72.

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.

Prentice, D.A., & Miller, D.T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160–164.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86, 638–641.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), The handbook of research synthesis (pp. 231–244). New York: Russell Sage Foundation.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276–1284.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335–392). Mahwah, NJ: Erlbaum.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115–129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 37–64). Mahwah, NJ: Erlbaum.

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.

Steiger, J.H., & Fouladi, R.T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. Behavior Research Methods, Instruments, and Computers, 4, 581–582.

Suter, G.W., II. (1996). Abuse of hypothesis testing statistics in ecological risk assessment. Human Ecological Risk Assessment, 2, 331–347.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26–30.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70, 80–93.

Thompson, B. (2002a). “Statistical,” “practical,” and “clinical”: How many kinds of significance do counselors need to consider? Journal of Counseling and Development, 80, 64–71.

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, 31, 24–31.

Thompson, B. (2004). The “significance” crisis in psychology and education. Journal of Socio-Economics, 33, 607–613.

Thompson, B. (2006a). Foundations of behavioral statistics: An insight-based approach. New York: Guilford.

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

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

Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? Journal of Experimental Education, 75, 109–125.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594–604.

Yoccuz, N.G. (1991). Commentary: Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecology Society of America, 72, 106–111.

Ziliak, S.T., & McCloskey, D.N. (2004). Size matters: The standard error of regressions in the *American Economic Review*. Journal of Socio-Economics, 33, 527–546.