What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes

by Bruce Thompson

An improved quantitative science would emphasize the use of confidence intervals (CIs), and especially CIs for effect sizes. This article reviews some definitions and issues related to developing these intervals. Confidence intervals for effect sizes are especially valuable because they facilitate *meta-analytic thinking* and the interpretation of intervals via comparison with the effect intervals from related prior studies. Several recommendations for the thoughtful use of such CIs are presented.

Recent annual meetings of the American Educational Research Association (e.g., the 1998 and 1999 "Royal Rumbles" sponsored by the Educational Statisticians special interest group), the American Psychological Society, and the American Psychological Association (APA) featured formal debates between distinguished scholars over whether statistical significance tests should or should not be banned from journal articles (cf. Hunter, 1997). In 1996, following almost 2 years of deliberations, the APA Board of Scientific Affairs appointed its Task Force on Statistical Inference to make recommendations regarding a possible ban. In 1999 the Task Force issued its recommendations.

The APA Task Force did not recommend that statistical significance tests should be banned from journals. But the Task Force did recommend a number of reforms of contemporary analytic practices. Three of these recommendations are particularly relevant to the present discussion and have also influenced the recently released fifth edition of the APA *Publication Manual* (2001).

The Research News and Comment section publishes commentary and analyses on trends, policies, utilization, and controversies in educational research. Like the articles and reviews in the Features and Book Review sections of *ER*, this material does not necessarily reflect the views of AERA nor is it endorsed by the organization.

Three Recommendations

Report Effect Sizes

An effect size characterizes the degree to which sample results diverge from the null hypothesis (cf. Cohen, 1988, 1994). For example, if a researcher hypothesizes that the *SD*s of three populations are equal, and the sample *SD*s are all the same (regardless of what they are), then the effect size is zero. As the sample results increasingly diverge from whatever is specified by the null hypothesis, the effect size will increasingly diverge from zero.

The APA Task Force strongly urged authors to report effect sizes (e.g., Cohen's d, Glass' delta, η^2, or adjusted R^2). The Task Force did not recommend the use of any one effect size from among the several dozens of available choices (Elmore & Rotou, 2001; Kirk, 1996; Synder & Lawson, 1993; Thompson, 2002). But the Task Force emphasized, "*Always* provide some effect-size estimate when reporting a p value" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599, emphasis added), and noted "reporting and interpreting effect sizes in the context of previously reported effects is *essential* to good research" (p. 599, emphasis added).

The 1994 edition of the APA *Publication Manual* "encouraged" (p. 18) authors to report effect sizes. However, there are now 11 empirical studies (Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000) of 1 or 2 volumes of 23 different journals demonstrating that this encouragement was not effective, perhaps because only encouraging effect size reporting

> presents a self-canceling mixed-message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, 1999, p. 162)

Thus, editorial policies at 19 journals now formally require effect size reporting (cf. McLean & Kaufman, 2000; Murphy, 1997; Snyder, 2000), including two journals circulated to more than 50,000 persons (i.e., the flagship journals of the American Counseling Association and the Council for Exceptional Children). Additional journals have plans to announce similar requirements this year. Furthermore, in response to Task Force recommendations, the 2001 APA *Publication Manual* now notes 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. . . . The general principle to be followed . . . is to 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. (pp. 25–26, emphasis added)

Report Confidence Intervals

Second, the Task Force (1999) emphasized that confidence intervals (CIs) are very useful, and again emphasized the importance of interpreting results in a given study by explicit comparisons with related results in prior studies (p. 599). Regarding the use of CIs, the 2001 APA *Publication Manual* suggests that CIs represent "in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (p. 22).

Use Graphics

Third, the Task Force recommended the use of graphics to enhance the interpretation and communication of results. This emphasis complements CI reporting, because CIs are readily amenable to graphical presentation.

Purpose of This Article

The recommended interpretation of effect sizes and CIs leads quite naturally to a suggestion to report CIs for effect sizes, even though the Task Force and the 2001 *Publication Manual* did not directly address this third practice. The purpose of this article is to illustrate these applications by portraying what a future quantitative science employing these tools might look like, and indeed how much better and more exciting such a science might be.

Confidence Interval Definitions

A brief discussion of the definition of CIs might first be helpful. Some misconceptions regarding CIs should be expected, because although intervals have long been recommended for researchers (cf. Chandler, 1957), empirical studies of reporting practices show that CIs have been reported infrequently to date (Finch, Cumming, & Thomason, 2001; Kieffer, Reese, & Thompson, 2001). 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.

Some textbooks define a CI as

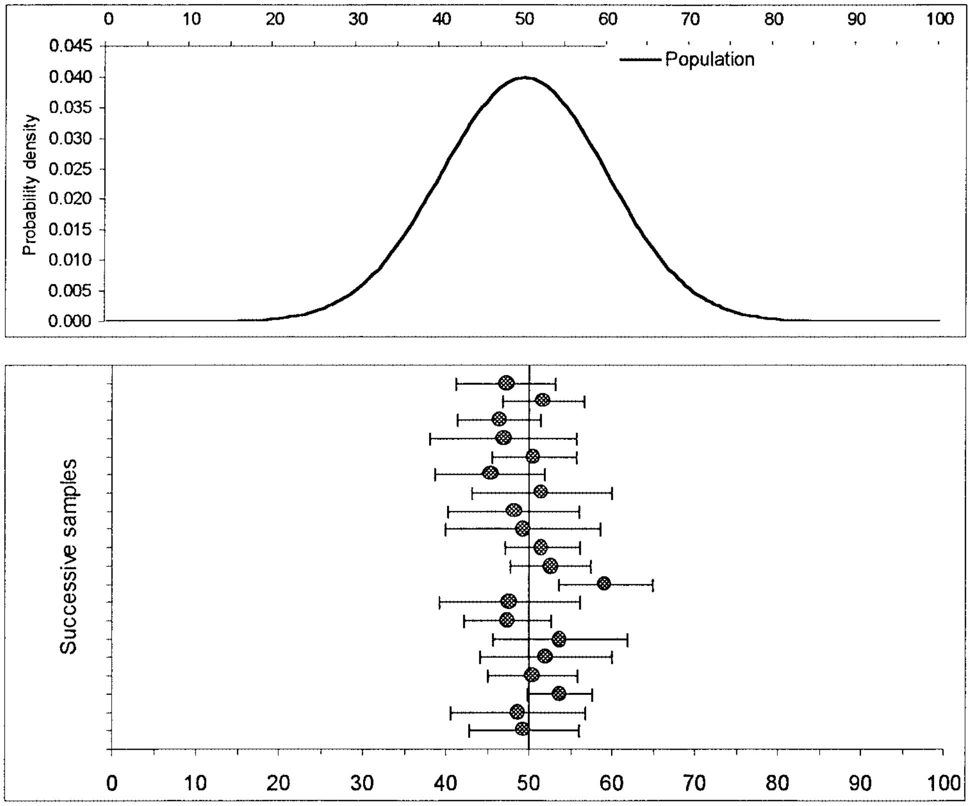
> . . . a range of numbers believed to include an unknown population parameter. Associated with the interval is a measure of the confidence we have that the interval does indeed contain the parameter of interest. (Aczel, 1995, p. 205)

But a better definition describes a given CI as one interval from among an infinite or at least large sample of CIs for a given parameter in which 1–α% of the intervals would capture the population parameter (cf. Moore & McCabe, 1993, p. 433; Tietjen, 1986, p. 35). As Hinkle, Wiersma, and Jurs (1998) explained using the example of a CI about an estimated mean,

> Theoretically, suppose we compute the sample means of all possible samples of size 20 and constructed the 95-percent confidence intervals for the population mean using all of these samples means. Then, 95-percent of these intervals would contain μ and 5 percent would *not.* (p. 222)

In short, both the parameter estimate and the endpoints for a single CI are influenced by sampling error, and so change from sample to sample.

Figure 1[1] illustrates these various dynamics. The figure presents a series of 20 random samples each of *n* = 12 cases from a population in which the mean is 50.0 and the *SD* is 10.0. In each of the 20 samples, the mean and its 95% confidence interval are both estimated. Several conclusions are readily drawn from Figure 1: (a) because of the sampling error variance, the estimated parameter (e.g., mean) varies some-



Note. The **top** portion of this figure presents the population of scores with a mean of 50.0 and a standard deviation of 10.0. The **bottom** portion of the figure presents the sample means (shaded circles) and the 95% CIs about each mean (bars) for 20 independent samples from the population.

FIGURE 1. *Population and a series of 20 estimated 95% intervals for sample (*n *= 12) means.*

what from sample to sample; and (b) the width of the interval (i.e., its precision) also varies somewhat from sample to sample (e.g., the 18th interval is considerably narrower than the 19th interval). But most relevant to the present discussion is the fact that one of the displayed intervals (12) did not capture the population mean. This reflects the expectation that 1 in 20 of a large number of 95% intervals, on the average, will not capture the population parameter (e.g., μ, ρ, σ). Cumming and Finch (2001) and others provide more elaboration of this point.

Note that a given interval either does or does not capture the parameter. This is a binary outcome with only these two discrete possibilities, just as one can only be pregnant or not pregnant, but cannot be 95% pregnant. Strictly speaking, it is not best practice to describe a single interval as 95% likely to capture a population parameter, because any 95% confidence statement applies, not to a single CI from a single sample, but to a large or infinite set of intervals constructed from a large or infinite number of samples. The confidence that one may vest in a large or infinite set of intervals capturing a population parameter does not guarantee that a single interval in a study actually did capture the population parameter being estimated.

It is also important to emphasize that, as for all statistical procedures, intervals for effect sizes will be incorrect if the assumptions for the computations (e.g., independence of observations, normality) are violated. Thus, these estimates are not a panacea applicable when assumptions are violated.

Intervals for Effect Sizes Versus Other Estimates

Confidence intervals about effect sizes are not the same as CIs for other parameter estimates, such as means. Nor are the effect size intervals computed in the same manner as CIs for statistics such as means or *SD*s. Confidence intervals for sample statistics, such as means and *SD*s, can be computed using formulas, and these computations have been incorporated in commonly used statistical programs for several decades.

Constructing CIs about effect sizes, on the other hand, raises two somewhat daunting technical difficulties. First, *noncentral t* and *F* distributions (cf. Pearson & Hartley, 1972), with which many researchers are unfamiliar, must be used to construct these effect size intervals (Fleishman, 1980; Steiger & Fouladi, 1997); these are not the *central t* and *F* distributions taught in most contemporary doctoral programs or statistics textbooks. Second, a generic formula cannot be employed to compute CIs for effect sizes, and instead computerized software estimation must be used to address each different research situation and result. Fortunately, commonly available software (e.g., SPSS) may be programmed to provide these so-called "iterative" estimates (Bird, 2002; Smithson, 2001). Iterative estimation is also necessary in various other statistical procedures with which many researchers may be more familiar, such as the estimation of communalities or of rotations in factor analysis.

A small heuristic data set may be used to illustrate the difference between computing an interval using a formula and estimating an interval using iteration (i.e., sequential estimation conducted until a statistical accuracy criterion is met). The "nil" null that $\mu = .0$ might be tested. Assume a sample data set with seven scores ranging in increments of .05 from .2 to .5. For these sample data $M = .35$ ($SD = .108$). Unlike the interval for an effect size, a formula can be used to compute the 95% CI about the mean, and the CI for the mean would be computed to be [.25 to .45]:

$$\begin{aligned} \text{CI}_{95\%} &= M \pm t_{(\alpha = .05,\, df = n-1)}\ [SD / \sqrt{n}] \\ &= .35 \pm 2.447\ [.108 / \sqrt{7}] \\ &= .35 \pm .10. \end{aligned}$$

This interval for the mean and the sample data is presented graphically in the upper portion of Figure 2.

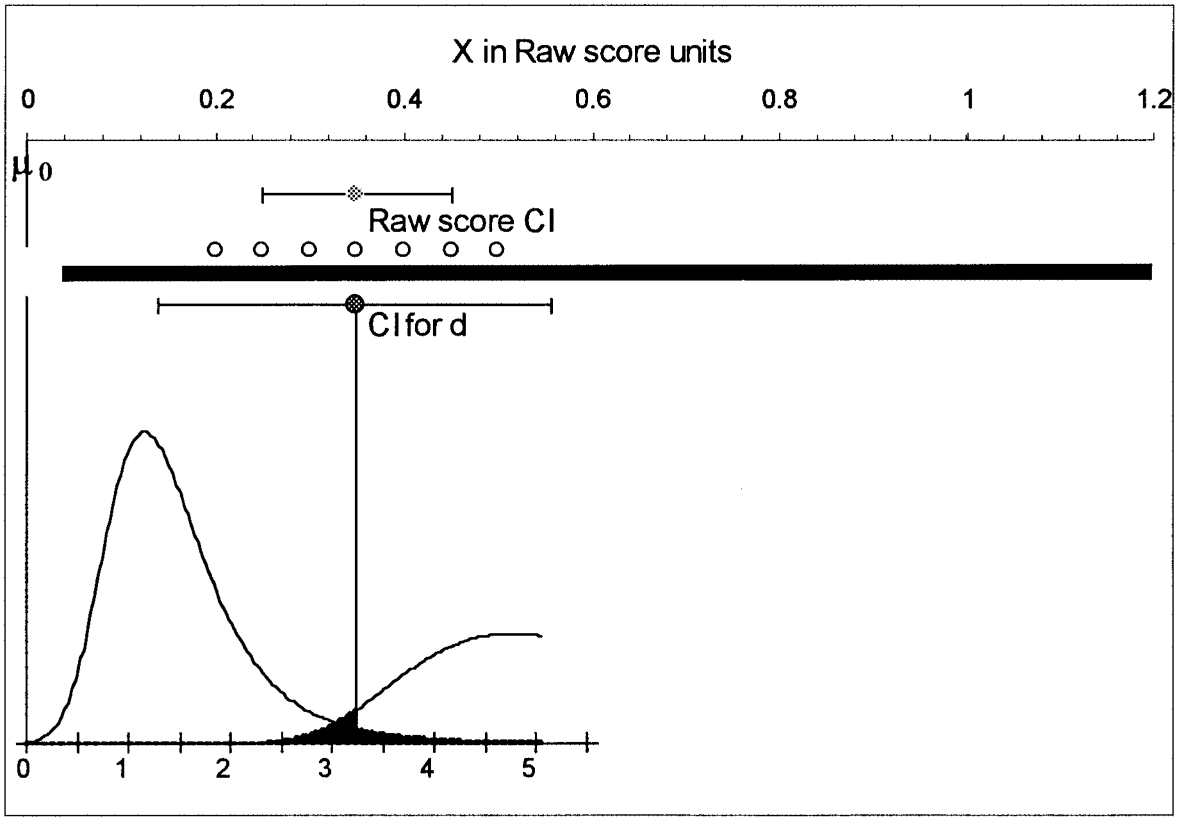
The *d* (cf. Cohen, 1988; Roberts & Henson, 2002) for these data equals 3.24 (.35 / .108). The confidence interval about this *d* can be computed using software such as that provided by Smithson (2001) or Cumming and Finch (2001). The result is presented in the lower portion of Figure 2. This interval is computed to be [1.29 to 5.15].

Figure 2 makes several important heuristic points. First, Figure 2 presents the CIs for both the mean and the related *d* effect size in a single figure, to emphasize that the scales or metrics for these two intervals are different. The scale of the CI for the mean, presented at the top of the figure, is in the metric of the original measurements. The scale for the standardized effect size (*d*) presented at the bottom of the figure, on the other hand, differs, because the scale of the original observations has been removed from this estimate by division by that metric. We use division by an *SD* to create standardized or *metric-free* effect sizes when the original measurements have a scale with no intrinsic meaning. We also use metric-free effect sizes when within a given literature we wish to compare effect sizes as "apples and apples" across different studies that used measures of different lengths or different scoring and thus involved *SD*s differing as a function of study design.

Second, Figure 2 also illustrates the two noncentral *t* distributions that must be employed in deriving the effect size interval. Note that these two noncentral distributions, presented at the bottom of the figure, at the boundaries of the CI for the effect size, unlike central *t*, are not symmetrical and are not centered at zero. Noncentral and central distributions differ more as (a) sample size is smaller or (b) effect size is larger.

A brief comment on the iterative use of noncentral distributions in building CIs about *d* is warranted, although this process is extremely technical and considerably more detail is available elsewhere (e.g., Cumming & Finch, 2001; Smithson, 2001). Because a formula cannot be used for this process, one tail of the effect size CI is iteratively estimated at a time. For example, as regards the left tail, and presuming a 95% interval is being constructed, a function of the noncentral distribution called the *noncentrality parameter* is estimated or guessed, and the percentage of the area under this curve that is immediately to the right of the *d* value or a function of this value is computed. The noncentrality parameter is iteratively tweaked until $\alpha/2$ (e.g., $.05/2 = .025$) of the area in the noncentral distribution is to the right of the *d* value or a function of this value. This area is shaded in Figure 2. Then the mean of this noncentral distribution is found, and that value defines the left boundary of the CI for *d*.

The process is then repeated to iteratively estimate the right CI boundary. Of course, the two boundaries can be found either left first or right first, because the boundaries are estimated independently using two different noncentral distribu-



Note. The **top** portion of the figure presents the scale for the observed raw score data, the mean (M = .35) and the 95% confidence interval (CI) for the mean, and the seven scores in the data set. The **bottom** portion of the figure presents the effect size, Cohen's d (i.e., d = 3.24), and the CI for this estimate.

FIGURE 2. *Ninety-five percent confidence intervals about* M *versus* d.

tions with different noncentrality parameters. As in factor analysis and other procedures where iteration is routinely used, the process of iteratively estimating CIs for effect sizes also can be automated within computer software, so that the iteration (although computationally demanding) is painless.

It is important to emphasize, as the figure also illustrates, that CIs about parameters such as means and effect sizes are not the same entities, even though the data and sample size are the same for both computations. This is illustrated in the fact that the widths or precisions of the intervals clearly differ for this example (i.e., the width of the interval for the mean is .2 [.45 – .25], but the width of the interval for the d effect size is 3.86 [5.15 – 1.29]).

Confidence Intervals Facilitate Meta-Analytic Thinking

As noted at the outset, the controversies regarding statistical significance testing have been extremely heated. 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). On the other hand, Abelson (1997) argued equally forcefully that if these tests did not exist, they would have to be invented. But there is one definitive area of agreement across the full spectrum of opinions regarding statistical significance testing: even scholars with diverse views of statistical significance testing increasingly emphasize the critical importance of replication in social science research (cf. Thompson, 1996; Robinson & Levin, 1997). This emphasis can be described as *thinking meta-analytically* (Cumming & Finch, 2001).

I define thinking meta-analytically as both (a) the prospective formulation of study expectations and design by explicitly invoking prior effect sizes and (b) the retrospective interpretation of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature. Confidence intervals about effect sizes are important because they facilitate exactly this sort of meta-analytic thinking.

The hypothetical results presented in Table 1 may be used to illustrate this process. The hypothetical example involved studies investigating whether high school students on medication to treat hyperactivity have mean verbal SAT scores that differ in either direction from a population mean of 500. The researcher found all relevant research and either computed effect sizes or located author reports of effect sizes. Of course, one day hopefully all authors will routinely report effect sizes to facilitate such literature summaries. In the example 10 studies had d values ranging from –.4 to +1.3, and sample sizes ranging from 4 to 53. The 95% intervals for these 10 effect sizes are presented in Figure 3.

The kind of meta-analytic thinking recommended here and by the APA Task Force would encourage the thoughtful integration of all prior related research when formulating expectations and designing a study. This is illustrated in Table 1 and

Table 1. ***Practical and Statistical Significance Statistics for 10 Previous Studies and One New Study***

	Results					
Study	*d*	*n*	*t* calc	*p* calc	Decision	CI for *d*
Prior Literature						
1	1.30	4	2.600	0.080	NS	−0.13 to 2.64
2	0.20	53	1.456	0.151	NS	−0.07 to 0.47
3	−0.40	34	−2.332	0.026	*	−0.75 to −0.05
4	0.50	17	2.062	0.056	NS	−0.01 to 1.00
5	0.70	9	2.100	0.069	NS	−0.05 to 1.42
6	0.65	11	2.156	0.056	NS	−0.02 to 1.29
7	0.80	7	2.117	0.079	NS	−0.09 to 1.64
8	0.60	12	2.078	0.062	NS	−0.03 to 1.21
9	0.40	25	2.000	0.057	NS	−0.01 to 0.80
10	0.30	35	1.775	0.085	NS	−0.04 to 0.64
Past research, pooled						
	0.278	*207*	*4.000*	*0.00009*	***	*0.14 to 0.42*
Current study						
11	0.45	19	1.962	0.065	NS	−0.03 to 0.92
Past and current, pooled						
	0.292	*226*	*4.390*	*0.00002*	***	*0.16 to 0.42*

Note. Pooled results are presented in italics. For the one-group case *t* = *d* (square root of *n*). Exact *p* calculated values can be found in Excel by using the function "TDIST(*t*, *n*−1, 2)". The weighted average effect size can be computed, for example for the 10 prior studies, as [(1.30 × 4) + (0.20 × 53) + …(0.30 × 35)] / [4 + 53 + …35] = [5.2 + 10.6 + …10.5] / 207 = 57.45 / 207 = .278. calc = calculated.

* *p* < .05 *** *p* < .001

Figure 3 by the synthesis of the prior effect sizes into a weighted average with $\overline{d}$ = .278 (*n* = 207).

It is important to emphasize that this empirical integration of effect sizes across studies presumes that the prior studies have all been screened to insure that each study is of acceptable quality. As is true throughout the process of inquiry, researchers also must invoke thoughtful judgment when integrating prior results (e.g., rather than "vote counting" how many prior studies achieved statistical significance or rotely integrating effect sizes even from poorly executed studies). For example, researchers might give more weight to studies deemed to have been better in quality (e.g., involved better measurement or had better checks for fidelity to an intervention protocol).

The researcher must also thoughtfully judge whether the studies being integrated are in some way biased. For example, reported effects might be unrepresentative if the "file drawer" problem (Rosenthal, 1979) meant that only statistically significant results were reported. Or published effect sizes might be positively biased if only studies with large effect sizes were admitted into the literature. Some effect sizes also may be statistically biased as a function of study design, although statistical corrections for these biases can sometimes be invoked to minimize these distortions (Kirk, 1996; Thompson, 2002). We are still learning which of the several dozen effect sizes are statistically biased and require correction, and which corrections are most appropriate (Roberts & Henson, 2002).

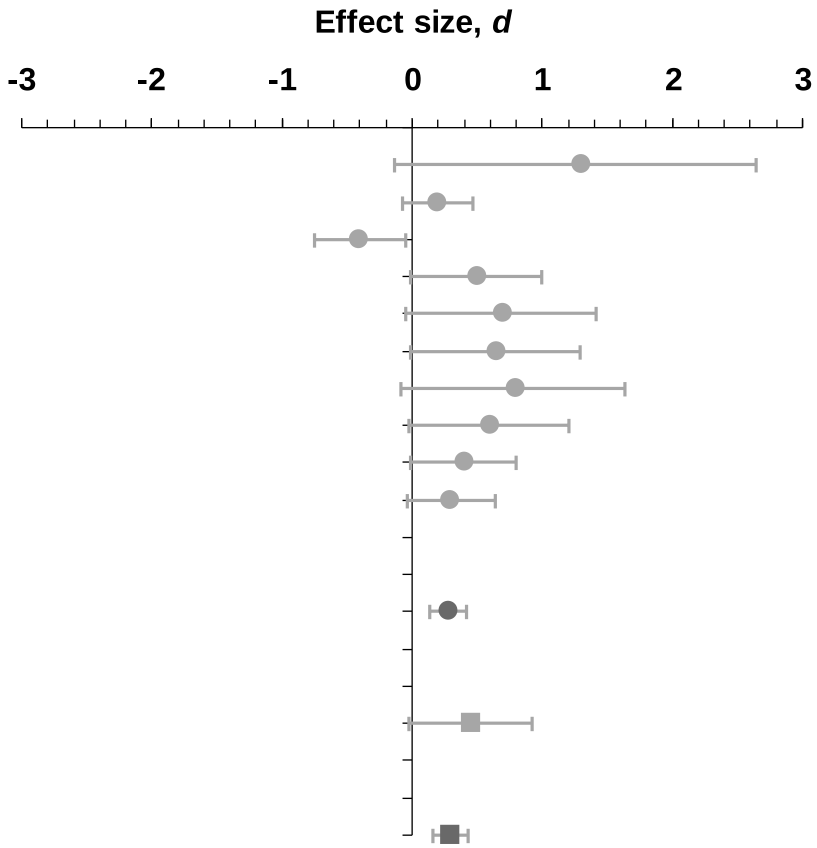
In my example the researcher then conducted a new study with *n* = 19 and found that *d* = +.45. The 95% CI for this effect size is reported in Figure 3. As a last step, as part of result interpretation, the researcher integrates this finding into the prior literature by computing a new, pooled CI for effect size presented at the very bottom of the figure.

Discussion

Some researchers erroneously equate CIs and statistical significance tests (Hagen, 1997; Knapp & Sawilowsky, 2001). Some confusion may arise because if a CI does not subsume zero, then a classical statistical significance test of the same data would always be statistically significant. But the most informative use of intervals does *not* evaluate whether a given interval subsumes zero, but instead compares intervals across studies. The distinction between significance testing and interval interpretation becomes clear when we realize that statistical tests may not be conducted absent a null hypothesis, but CIs can be constructed without a hypothesis. Confidence intervals can also be interpreted without hypotheses by couching interpretations in terms of the CIs from prior, related studies. The appeal of intervals is that across studies using intervals will ultimately lead us to the correct population value, even if our initial expectations are wildly wrong (Schmidt, 1996).

How Big is Big Enough?

One issue that confronts any researcher using either effect sizes or CIs about effect sizes is the question of what is a noteworthy effect. The field has not reached a definitive view on this matter. In his various books on power analysis (cf. Cohen, 1988), although he had reservations about doing so, Cohen proposed some tentative benchmarks for what might be deemed *small, medium,* and *large* effects as regards *d*. Because many of the several dozen effect sizes (e.g., η^2, ω^2) can be converted in *d*,



Note. The top 10 effect size intervals are for the studies from the prior literature, with the *d*'s represented as gray circles within the intervals. The pooled weighted average *d* (i.e., .278) is presented next, as a darker oval. The result in the present study ($d = .45$) is then presented as a gray square. Finally, the weighted average *d* across all 11 studies is presented last as a dark square.

FIGURE 3. *Integration of confidence intervals across studies.*

or vice versa, his tentative benchmarks potentially have wide applicability (Kirk, 1996; Thompson, 2002).

Cohen's hesitancy in presenting criteria for effect noteworthiness stemmed from the important admonition that the noteworthiness of an effect turns largely on what one is studying. Small but replicable effects for very important outcomes may be very noteworthy; extremely large effects may be needed for results to be noteworthy for relatively unimportant outcomes. For example, Gage (1978) pointed out that even though the relationship between cigarette smoking and lung cancer is relatively small (i.e., $\eta^2 = 1\%$ to 2%),

> Sometimes even very weak relationships can be important . . . [O]n the basis of such correlations, important public health policy has been made and millions of people have changed strong habits. (p. 21)

Of course, this has also occurred because the effect involving a very cherished outcome variable has been replicated numerous times.

Some people invoke Cohen's benchmarks for effects with more rigidity than he may have liked. As Thompson (2001) recently noted, "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" (pp. 82–83).

Researchers in the past should not have employed $p < .05$ as an atavistic escape from explicitly arguing why their results were noteworthy (Thompson, 1993). By the same token, the existence of effect size benchmarks should not justify abrogating the responsibility for arguing effect import in the specific context of a given study. It is not necessary to have universal benchmarks regarding what effect sizes may be deemed noteworthy. The reader with a value system widely different than that of an author might reasonably disagree with the author about whether the effect size is noteworthy and then simply ignore the study. But once reporting effect sizes becomes normatively standard practice, at least all authors and readers will then be in a position to evaluate how replicable or stable are the effects within a given area of inquiry.

Three Recommendations for Practice

Several recommendations for practice are suggested here:

1. *Report and explicitly interpret effect sizes in the context of effect sizes from prior related studies and not by invoking rigid benchmarks.* The potential benefits of reporting and interpreting an effect size (e.g., Cohen's *d*, Glass' delta, η^2, or adjusted R^2) arise not from interpreting effects against benchmarks, but rather by comparing effect sizes directly with the effects reported in related prior studies (Wilkinson & APA Task Force, 1999, p. 599). The overly rigid use of fixed benchmarks for small, medium, and large effects fails to consider the possibility that small, replicable effects involving important outcomes can be noteworthy, or that large effects involving trivial outcomes may not be particularly noteworthy.
2. *Report and interpret CIs, including intervals for effect sizes, in the context of related intervals from prior research, and not only by evaluating whether they subsume zero.* The potential benefits of computing CIs about either parameter estimates (e.g., means, *SD*s) or effect sizes (e.g., Cohen's *d*) do not arise from interpreting intervals by evaluating whether they subsume zero; that is merely null hypothesis significance testing in another guise (cf. Thompson, 1998). Rather, these benefits accrue by comparing intervals across related studies, thereby converging on population parameters.
3. *When literature reports more than a handful of CIs, consider summarizing all the intervals in a graphic form.* One potential benefit of reporting CIs is that intervals from prior and current studies can readily be presented in picture form similar to that used in Figure 3. Note that Figure 3 contains the information presented in Table 1, but the graphic presentation may be more digestible.

Various scholars have presented a thoughtful case for more frequent use of graphics in research, including Loftus (1993), Tufte (2001), and Wainer (2000). Put simply, in some cases a picture is worth a thousand words (or several tables). First, from Figure 3 at a glance one can determine that 10 of the 11 studies had positive effects, and that the weighted averages of the 10 prior or of all 11 studies (past plus present) were positive. From such a picture one could easily glean an impression of a large literature involving even 50 or 60 (or more) effect sizes. Second, the widths of the intervals also facilitate an impression regarding the precision of effect estimation across the studies within a literature.

An important way to think about a CI is as a representation of all reasonably plausible parameter estimates based on results in either a single study or in a literature. For example, in Study 1 $d = 1.3$; but based on the study's CI, d might also be estimated as −.13, .0, 1.0, 2.6, or any other number between −.13 and 2.64. A graphic showing wide intervals across all or most studies indicates that too many studies are being poorly conducted because precision is so limited. It is conceivable that researchers may not recognize insufficient precision in literature until they see the big picture (perhaps literally) as regards such interval widths across studies.

Report Effect Sizes Even for Nonsignificant Results

Some have argued that effect sizes should only be reported when results are statistically significant (Robinson & Levin, 1997) to minimize the likelihood that researchers will overinterpret results that are actually nonreplicable. This danger is minimized if the interpretative framework never focuses on the result in a single study in isolation, but rather *always* focuses on the result in a given study in direct comparison with effect sizes in related prior studies. The Figure 3 example makes it clear why presenting effect sizes even for statistically nonsignificant results is so important.

In the Figure 3 example only 1 of the 11 studies had statistically significant results ($\alpha = .05$): $d = -.40$, $p = .026$. This also happened to be the only 1 of the 11 studies in which the sample mean was less than the population mean verbal SAT score of 500. Note that if the file drawer problem (Rosenthal, 1979) had governed, researchers might wrongly conclude that students on hyperactivity medication tend to score less than 500 on the verbal SAT. Only this result would have been published, and everyone would have posited a pattern in the wrong direction.

Here the aggregation of effects across the previous studies and the current study (i.e., the pooled result presented at the bottom of Table 1 and graphically in Figure 3) suggests an overall pattern of positive effects, and the precision of the estimated effect size interval is fairly narrow, as reflected in the narrow width of the aggregated CI. The result also illustrates why Rosnow and Rosenthal (1989) wrote that "surely, God loves the .06 [level of statistical significance] nearly as much as the .05" (p. 1277); the pooling of studies each with $p = .06$ can indeed lead to the discovery of a replicable result with important practical consequences.

Summary

A brighter day is dawning in which researchers will ask not only if a sample result is likely but also if an effect is practically noteworthy or replicable (Kirk, 1996, 2001). The incremental progress toward creating a better science would be facilitated by reporting and interpreting effect sizes, CIs for relevant sample results (e.g., means), and CIs about effect sizes.

Short and accessible treatments of effect size choices have been provided by Kirk (in press), Olejnik and Algina (2000), Snyder and Lawson (1993), and Thompson (2002). Software for computing CIs for various effect sizes has been reported by Bird (2002), Cumming and Finch (2001), Smithson (2001), and Steiger and Fouladi (1992). Some of these programs can analyze raw data while others use summary statistics; some programs run under Excel (e.g., Cumming & Finch, 2001) or SPSS (e.g., Smithson, 2001) or stand-alone (e.g., Steiger & Fouladi, 1992). And in many cases these applications can be downloaded off the Web for free or for only a nominal charge.

Thinking meta-analytically itself, even absent other needed improvements in contemporary practices, would result in an improved science of discovery. If educational research can make a difference in the lives of children and other stakeholders, as we often presume when conducting inquiry, we ought to report and interpret our results in ways that do everything possible to optimize the potentials for insight and impact.

NOTES

Bruce Thompson is Professor and Distinguished Research Fellow, Department of Educational Psychology, Texas A & M University, College Station, TX 77843-4225; the author receives e-mail and provides related reprints at http://www.coe.tamu.edu/~bthompson. His specializations include statistics, measurement, and program evaluation.

[1] Figure 1 and the other figures reported here were constructed using the Exploratory Software for Confidence Intervals (ESCI) developed by Geoff Cumming at La Trobe University in Australia. This software runs under Excel, is a wonderful tool for both learning about and teaching important effect size and CI concepts, and can be downloaded from http://www.latrobe.edu.au/psy/esci.

REFERENCES

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Aczel, A. D. (1995). *Statistics: Concepts and applications.* Chicago: Irwin.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

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

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197–226.

Chandler, R. E. (1957). The statistical concepts of confidence and significance. *Psychological Bulletin, 54*, 429–430.

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*, 532–575.

Elmore, P., & Rotou, O. (2001, April). *A primer on basic effect size concepts.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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.

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, *40*, 659–670.

Gage, N. L. (1978). *The scientific basis of the art of teaching*. New York: Teachers College Press.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: Houghton Mifflin.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3–7.

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. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213–218.

Kirk, R. E. (in press). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology*. Oxford, UK: Blackwell.

Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, *70*, 65–79.

Loftus, G. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments and Computers*, *25*, 250–256.

McLean, J. E., & Kaufman, A. S. (2000). Editorial: Statistical significance testing and other changes to *Research in the Schools*. *Research in the Schools*, *7*(2), 1–2.

Moore, D. S., & McCabe, G. P. (1993). *Introduction to the practice of statistics* (2nd ed.). New York: W.H. Freeman.

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

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

Pearson, E. S., & Hartley, H. O. (1972). *Biometrika tables for statisticians* (Vol. 2). Cambridge, UK: Cambridge University Press.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement, 62,* 241–253.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21–26.

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

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

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the 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.

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention*, *23*, 145–150.

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

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.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361–377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*, 799–800.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*, *11*, 157–169.

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

Tietjen, G. L. (1986). *A topical dictionary of statistics*. New York: Chapman and Hall.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology*, *10*, 413–425.

Wainer, H. (2000). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Mahwah, NJ: Erlbaum.

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

Manuscript received September 14, 2001
Revisions received January 15, 2002
Accepted January 23, 2002