Journal of Educational Psychology

A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology

Shuyan Sun, Wei Pan, and Lihshing Leigh Wang

Online First Publication, September 20, 2010. doi: 10.1037/a0019507

Journal of Educational Psychology

0022-0663/10/$12.00 DOI: 10.1037/a0019507

A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology

Shuyan Sun, Wei Pan, and Lihshing Leigh Wang
University of Cincinnati

Null hypothesis significance testing has dominated quantitative research in education and psychology. However, the statistical significance of a test as indicated by a *p*-value does not speak to the practical significance of the study. Thus, reporting effect size to supplement *p*-value is highly recommended by scholars, journal editors, and academic associations. As a measure of practical significance, effect size quantifies the size of mean differences or strength of associations and directly answers the research questions. Furthermore, a comparison of effect sizes across studies facilitates meta-analytic assessment of the effect size and accumulation of knowledge. In the current comprehensive review, we investigated the most recent effect size reporting and interpreting practices in 1,243 articles published in 14 academic journals from 2005 to 2007. Overall, 49% of the articles reported effect size—57% of which interpreted effect size. As an empirical study for the sake of good research methodology in education and psychology, in the present study we provide an illustrative example of reporting and interpreting effect size in a published study. Furthermore, a 7-step guideline for quantitative researchers is also summarized along with some recommended resources on how to understand and interpret effect size.

Keywords: NHST, effect size, statistical significance, practical significance, confidence intervals

Effect size measures as a criterion for practical significance has been recommended for a long time to supplement null hypothesis significance testing (NHST) to get better statistics and results (e.g., American Educational Research Association [AERA], 2006; American Psychological Association [APA], 2001; Anderson, Burnham, & Thompson, 2000; Kirk, 1996; Plucker, 1997; Robinson & Levin, 1997; Thompson, 1998b, 1999c, 1999d; Thompson & Snyder, 1997). The effectiveness of this recommendation is worthy of a methodological review. Thus, the purpose of the present study was to investigate the effect size reporting and interpreting practices in academic journals in education and psychology areas. To build the present study into a solid theoretical framework, we review the problems and misuses of NHST, and we emphasize the importance of effect size and its relationship with NHST, confidence intervals, and meta-analytical assessment of effect size. An illustrative example of reporting and interpreting effect size in a published study is also presented. To facilitate quantitative researchers' use of effect size, we recommend resources on how to understand and use effect size, and we provide a seven-step guideline for quantitative researchers to conclude the study.

NHST

NHST is a traditional and popular approach to make statistical inference about research questions (Anderson et al., 2000). It is considered to be an objective, scientific procedure for knowledge accumulation (Kirk, 1996). It frames research questions in terms of two contrasting statistical hypotheses. For instance, when the purpose is to examine the effect of a treatment, the null hypothesis states that "the experimental group and the control group are not different with respect to [a specified property of interest] and that any difference found between their means is due to sampling fluctuation" (Carver, 1978, p. 381), whereas the alternative hypothesis states the opposite for a two-tailed test of mean difference as population parameter. When a population correlation is under investigation, null hypothesis states that there is no correlation between two variables, whereas the alternative hypothesis states that there is a correlation between them. Same patterns are applicable to two-tailed hypotheses for other population parameters. One-tailed hypotheses are formulated in a similar way with a special feature that alternative hypothesis states the direction of the prediction, and the null hypothesis states the opposite. With hypotheses stated, applying an appropriate statistical model yields a *p*-value, an observed probability that the statistical test would have yielded a statistic equal or greater than the one obtained, if the samples used had been drawn randomly from the same population that characterizes the null state. An alpha, a designated significance level, acts as a decision criterion, and the null hypothesis is rejected only if the *p*-value yielded by the test is smaller than the value of the alpha.

Purpose of NHST

The purpose of NHST is to provide a framework for making inference from a sample to the population in the face of uncertainty caused by sampling error (Kline, 2004). NHST addresses whether observed effects or relations stand out above sampling error by test statistic and its *p*-value, though it is not as useful for estimating the

Shuyan Sun, Wei Pan, and Lihshing Leigh Wang, School of Education, University of Cincinnati.

Correspondence concerning this article should be addressed to Shuyan Sun, Educational Studies Program, Dyer Hall 409A, University of Cincinnati, P.O. Box 210049, Cincinnati, OH 45221. E-mail: sunsn@mail.uc.edu

magnitude of these effects (Chow, 1996). The *p*-values estimate the probability of sample results deviating as much or more than do the actual sample results from those specified by the null hypothesis (Cohen, 1994; Kirk, 2001). The final outcome of NHST is the decision to reject or fail to reject null hypotheses; it meets the needs of research questions that do require a dichotomous answer. For instance, researchers in some fields—such as engineering, business management, and environmental studies—can utilize NHST to estimate the costs of different decisions in dollars, life expectancy, or some other quantitative and objective metrics. The expected gains and losses are evaluated to select the best action from several well-defined alternatives in the face of uncertainty. This approach is well known as statistical decision theory (Kline, 2004). Unfortunately, it is usually not possible in social and behavioral research.

Problems With NHST

Though NHST was considered to be an objective, scientific procedure for knowledge accumulation (Kirk, 1996), it holds a controversial status in social and behavioral research: On the one hand, it is an integral part of scientific research; on the other hand, it has been surrounded by controversy and criticisms (Kirk, 1996; Robinson & Wainer, 2002). The earliest serious challenges to NHST dated back to 1938 when Joseph Berkson published his article to question the logic and usefulness of NHST (Berkson, 1938). Since then, criticisms of NHST have noticeably intensified (e.g., Anderson et al., 2000; Carver, 1978; Cohen, 1990, 1994; Falk & Greenbaum, 1995; Harlow, Mulaik, & Steiger, 1997; Henson & Smith, 2000; Katzer & Sodt, 1973; Kirk, 1996; Robinson & Wainer, 2002; Schmidt, 1996; Yates, 1951).

The fundamental problem with NHST is not that it is methodologically wrong; the misuse of NHST is the fault. NHST is built in the framework of sampling distribution of statistics and addresses the issue of sampling error in the process of estimating population parameters. Its null hypotheses are sometimes appropriate; and it meets the needs of dichotomous decision very well in the hard science, such as engineering, though it is usually not possible in social behavioral science (Kline, 2004). However, the interpretation and application of the results from NHST are often problematic (Anderson et al., 2000; Ives, 2003).

The problems of NHST can be summarized into the following three aspects. First, the NHST procedure does not tell researchers what they want to know. In other words, NHST and scientific inference address different questions. Researchers are interested in the probability of the null hypothesis being true given the data collected. However, NHST estimates the probability of obtaining the data given null hypothesis is true, a logic that is a reverse of research logic. Therefore, successful rejection of the null hypothesis cannot be interpreted that the theory that guides the test is affirmed. As Cohen (1994) observed, a statistical significant test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997). Associated with this illusion are incorrect widespread beliefs that a *p*-value measures the likelihood of sampling error, replication, and whether hypotheses are true (Kirk, 1996; Kline, 2004). As Kline (2004) pointed out, these false beliefs may not be solely NHST users' fault; the logical underpinnings of contemporary NHST are not entirely consistent.

The second problem is that by adopting an arbitrarily fixed level of significance alpha, researchers turn a continuum of uncertainty into an artificial dichotomous reject-or-do-not-reject decision. There is no theoretical basis for the choice of alpha except for the conventional values; thus, decision based on this significance level is practically meaningless (Anderson et al., 2000). This decision strategy can lead to the situation in which two researchers obtain identical treatment effects but draw different conclusions from their research (Cohen, 1994; Kirk, 1996; Thompson, 1997; Young, 1993). The practical difference between a *p*-value of .049 as opposed to one of .051 is certainly not as dramatic as the dichotomous decision based on conventional choices of alpha level .05. Because of this dichotomous decision rule, the failure to reject null hypothesis may be mistakenly interpreted as evidence for accepting the null (Kirk, 1996). Moreover, it does not directly tell researchers the size of an effect or whether there is any theoretical, practical, or clinical importance (Chow, 1988; Kirk, 1996; Shaver, 1993).

The third problem is that null hypotheses are always false on a prior ground in the real world (Cohen, 1990, 1994; Johnson, 1995; Kirk, 1996; Kline, 2004). This problem is caused by an almost universal misuse of null hypothesis that null hypothesis means nil or zero (Cohen, 1994; Thompson, 1999c). It is very unlikely that the value of any population parameter is exactly zero, especially when zero implies the complete absence of an effect or association. For instance, the effects of two treatments are always different in some decimal places, thus asking whether they are different is a trivial exercise (Kirk, 1996). When nil hypotheses are rarely true on a prior ground, a decision to reject a nil null hypothesis simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. Increased sample size will eventually yield a statistical significance only if the null hypothesis is false (Biskin, 1998). Some scholars questioned whether inference could be extended from a theoretical population to actual sample values; in practice, the null hypothesis is essentially false, and therefore, statistical significance testing becomes a vain effort of demonstrating what is already known (Kirk, 1996; Thompson, 1993; Vacha-Haase & Thompson, 2004). Researchers do not always have to test nil null hypotheses (Thompson, 1997), but most researchers do so because this is what most computer packages assume, and it is more complicated to apply non-nil null hypotheses in many designs (Dar, Serlin, & Omer, 1994). The mindless use of the nil hypotheses further obviates the necessity of estimating expected effect sizes from prior literature as part of study design (Thompson, 1999a). Thus, NHST provides little information of scientific interest and, in this respect, is of little practical use in the advancement of knowledge. Therefore, the focus of research should always be proper estimation of the size of the treatment effect (Anderson et al., 2000); size of effect and whether an effect is replicable are what researchers really want and need to know (Kline, 2004).

These criticisms of NHST have lead researchers to explore alternative methods that can make data analysis more meaningful in the context of research problems. Though some authors (e.g., Carver, 1978; Schmidt & Hunter, 1995) have recommended complete elimination of significance testing, most scholars and professional organizations suggest that NHST should be supplemented with or placed in the context of additional information, such as confidence intervals and effect size (AERA, 2006; Ander-

son et al., 2000; APA, 2001, 2010; Cumming & Finch, 2002; Fan, 2001; Kirk, 1996, 2001; McLean & Ernest, 1998; Snyder & Lawson, 1993; Thompson, 1996, 1997, 2002b; Vacha-Haase & Thompson, 2004; Vaske, Gliner, & Morgan, 2002; Wilkinson & APA Task Force on Statistical Inference, 1999). Reporting effect size is probably the most frequent recommendation (Ives, 2003).

Effect Size

The Definition of Effect Size

Effect size can be broadly defined as any statistic that quantifies the degree to which sample results diverge from the expectations specified in the null hypothesis (Cohen, 1994; Kline, 2004; Thompson, 1998b, 2002b, 2008; Vacha-Haase & Thompson, 2004). There are dozens of effect size measures available, each with relative strengths and weaknesses for particular purposes (Henson, 2006; Kirk, 1996). The big family of effect size measures has been categorized into two broad groups: measures of mean differences and measures of strength of relations. The former is based on the standardized group mean difference and is represented by Cohen's *d*, Glass's Δ, and Hedges's *g* (Cohen, 1988; Glass, 1976; Hedges, 1981); the latter is based on the proportion of variance accounted for or correlation between two variables and is represented by *R*-squared (R^2) and eta-squared (η^2) (Kirk, 1996; Maxwell & Delaney, 1990; Snyder & Lawson, 1993). It should be noted that some effect size measures do not fall neatly into these two categories (e.g., the *I* index for hit rate by Huberty & Lowman, 2000).

The Importance of Effect Size

Effect size may be useful in at least three practical applications. First, before a study is carried out, estimates of anticipated effect sizes can be used to project the sample size that would be adequate for detecting statistically significant results. Minimum sample size that is adequate to detect a particular effect size can be calculated after estimating or selecting the values of the effect, alpha, and power. It will help reduce the risk of statistically nonsignificant results because of inadequate sample size (Olejnik, 1984; Plucker, 1997). Second, it enables researchers to inform judgment about the practical significance of the study, given the substantive context of the study (Kirk, 1996; Thompson, 2008). As Fan (2001) argued, *p*-value and effect size are two sides of one coin: They complement each other, but they do not substitute for each other; therefore, researchers should consider both sides. The purpose of research should be to measure the magnitude of an effect rather than simply its statistical significance (Cohen, 1990); thus, reporting and interpreting effect size is crucial. Third, because effect sizes are intended to be metric-free measures of the size of mean differences or the strength of relations, they may be used to compare the results of different studies with one another and to evaluate the replicability of results. If effect sizes are stable across studies or even generalizable over some variations in design or analysis, the results are replicable (Thompson, 2008). That is, they provide a statistical tool for meta-analysis that quantitatively synthesizes the effects across different studies. Explicitly reporting effect sizes helps meta-analysts avoid computing approximate effect sizes based on sometimes tenuous statistical assumptions and, therefore, more easily and more accurately synthesize findings across studies (Thompson, 1999a).

The Interpretation of Effect Size

Reporting effect size only is not enough; researchers should interpret and evaluate effect size for its practical significance (Kline, 2004). Thus, how to interpret effect size is also a crucial question. The common practice in interpreting effect sizes is to use the benchmarks for "small," "medium," and "large" effects offered by Cohen (1988). However, this is an unfortunate practice in that Cohen's benchmarks are not generally useful (Thompson, 1999a, 2008). Cohen offered these benchmarks as general guidelines for researchers working in unexplored territory "because they were needed in research climate characterized by a neglect of attention to issues of [effect size] magnitude" (Cohen, 1988, p. 532). In a relatively established area of research, it is inappropriate to apply Cohen's guidelines blindly (Glass, McGaw, & Smith, 1981; Thompson, 2008). The appropriate interpretation of effect size should focus on explicitly and directly comparing between effect sizes in new results and prior effect sizes in the related literature (Thompson, 2008); both the size and nature of the effect should be included in the interpretation (Henson, 2006). Effect sizes can inform practical significance, but they are not inherently meaningful. The importance and meaning of an effect size depend on multiple factors, such as the context of the study, the importance of the outcomes, and the size and nature of effect obtained in prior studies (Henson, 2006). Therefore, researchers should interpret the effect size both within and between studies.

Discrepancy Between *p*-Value and Effect Size

A test result that is statistically significant as judged by the *p*-value is not necessarily practically significant as judged by the effect size. Thus, a small *p*-value cannot be interpreted as the presence of an acceptable effect. There are four possible outcomes in a test: (a) a statistically significant *p*-value with a practically significant effect size, (b) a statistically nonsignificant *p*-value with a practically nonsignificant effect size, (c) a statistically significant *p*-value with a practically nonsignificant effect size, and (d) a statistically nonsignificant *p*-value with a practically significant effect size. There is no interpretational problem with the first two situations in that the consistency between *p*-value and effect size is achieved. Situation c could mistake a statistical significance for a practical significance, whereas Situation d fails to identify the practical significance of a statistically nonsignificant result (Rosenthal, Rosnow, & Rubin, 2000). Therefore, these latter two situations are considered as discrepancy between *p*-value and effect size, which would suggest possible threats to the design validity of the study.

Consequences of Not Reporting Effect Size

As Zientek, Capraro, and Capraro (2008) pointed out, "Not reporting effect size can be detrimental" (p. 212). Because of the potential discrepancy between *p*-value and effect size addressed in the previous section, reporting effect size becomes extremely important for both statistically significant and nonsignificant tests. A small *p*-value does not necessarily indicate a practical signifi-

cance of the effect; on the other hand, a large *p*-value for a practically significant effect may be due to the limited statistical power to detect such an effect. Even when all effects within single studies from the literature are statistically nonsignificant, a quite impressive, meta-analytically pooled effect may arise (Thompson, 2007) by examining the overlap of confidence intervals of the effect sizes across the studies. Thus, researchers should always report effect size for both significant and nonsignificant tests; it contextualizes the impact of the study directly and explicitly (Thompson, 2007). If effect size is not reported in primary analysis, researchers who are interested in doing secondary analysis or meta-analysis have to use an approximate conversion formula to estimate the effect size because of no access to the raw data (Grissom & Kim, 2005); therefore, researchers working with primary data should report exact effect size to improve the accuracy of estimation for secondary analysis. To sum up, not reporting effect size is detrimental not only to a single study but also to the knowledge accumulation in the long run.

Confidence Intervals and Effect Size

Reporting confidence intervals as an alternative to *p*-value in NHST is also frequently recommended (e.g., Dar et al., 1994; Meehl, 1997; Schmidt, 1996; Serlin, 1993). A $(1 - \alpha)\%$ confidence interval for a statistic yields a pair of statistics that, over repeated samples, includes the parameter with a probability of $1 - \alpha$ (Steiger & Fouladi, 1997). As an interval estimator, a confidence interval has two functions: (a) Confidence intervals can be used to indicate the precision of the estimate of the parameter, such as population mean or population standard deviation—the smaller the confidence interval is, the more precise the estimation—and (b) the parameter can emerge across studies as the overlaps of confidence intervals converging on the same parameter (Thompson, 1998a, 1999a, 2007; Wilkinson & APA Task Force on Statistical Inference, 1999; Zientek et al., 2008), even when all studies investigating the same research questions make erroneous estimates of the parameters (Schmidt, 1996; Thompson, 1999a).

It is worthy to note that the second function of confidence interval is often misinterpreted as reflecting the certainty that a confidence interval captures the true parameter (Thompson, 2007). The endpoints of a confidence interval are random variables estimated on the basis of sample data (Thompson, 1999a); therefore, a confidence interval constructed from a single sample is just one of the numerous possible estimates and, thus, does not indicate the chance of including parameter in the interval (Falk & Greenbaum, 1995). Computing 95% confidence intervals for a statistic means that if infinitely many random samples were taken from the same population, exactly 95% of the confidence intervals would capture the parameter, and exactly 5% would not (Thompson, 2007). Furthermore, blindly interpreting confidence intervals only against the standard of whether zero point is included is nothing more than rejecting or not rejecting on the basis of *p*-value (Cortina & Dunlap, 1997; Thompson, 1999a, 2008).

Constructing confidence intervals is a very important tool of result description, and it can be constructed without conducting the hypothesis testing (M. M. Capraro, 2005; Zientek et al., 2008). Though "confidence intervals and NHST calculations are based on precisely the same information" (Cortina & Dunlap, 1997, p. 170), confidence interval includes the parameter estimates that would not lead to the rejection of null hypothesis (Yetkiner, Capraro, Zientek, & Thompson, 2008). A confidence interval offers a range of values, which are not included in NHST, as the estimates of the population parameter; thus, constructing confidence interval is more informative than NHST.

It should be noted that as statistics, effect size measures have their limitations and are not a panacea for all the problems in NHST and result interpretation. First, effect size estimation depends on means and standard deviations of the sample and, thus, will vary from sample to sample; therefore, any single effect should be interpreted with caution. Second, as a point estimator of population effect, effect size does not indicate the accuracy of the estimation. As proposed by Henson (2006), the limitations of effect size can be overcome in two ways: (a) meta-analyzing effect sizes across studies because the true parameter value can emerge by comparing effect sizes over enough replications, and (b) similar to constructing confidence intervals for population mean, constructing confidence intervals for the effect size within individual studies to include all values that are reasonably expected instead of using one point to estimate the true population effect. These two approaches are also recommended by Kline (2004) and Thompson (2002b, 2007, 2008).

Effect Size, Replication, and Meta-Analytic Assessment

In social science research, replication has not been given as much attention as in the natural science (Henson, 2006; Kline, 2004). The generalizability of a single study is very limited when nonrandom sampling, inadequate sample size, common internal and external threats to validity, and possible violation of statistical assumptions are considered. Meta-analytic thinking or assessment would considerably facilitate knowledge accumulation in the field (Cumming & Finch, 2001; Henson, 2006). Meta-analytic assessment does not overemphasize the outcomes of statistical tests in individual studies (Kline, 2004); instead, it emphasizes the need to explicitly design and place the studies in the context of the effects of prior research, and the reporting and interpreting effect size provide a vehicle to make comparisons across studies more explicit (Henson, 2006). It is particular valuable when the formulation of effect sizes can be guided by findings in previous studies before the study is conducted, and comparison of effect sizes across studies can be made after the study is completed (Thompson, 2002a, 2002b). Constructing confidence intervals for effect size can facilitate meta-analytic assessment of effect size and advance scientific inquiry; researchers do not have to conduct a formal and complete meta-analysis study as proposed by Hedges and Olkin (1985), but they should develop the capacity to meta-analytically assess their own research by comparing their results with previous findings and examining the overlaps of confidence intervals of the effect size across the studies (Thompson, 2002b).

Changing Publication Policies

In response to the criticisms about NHST and to raise awareness of the importance of effect size, some journals and academic associations have changed their publication policies. In 1994, *Educational and Psychological Measurement*, the first journal that required effect size reporting, published its editorial requirements (Thompson, 1994). After that, more and more journals definitively

require effect size reporting in their publication policies. Currently, there are at least 24 journals that have such a policy in place (for a list of these journals, visit Bruce Thompson's homepage at http://www.coe.tamu.edu/~bthompson/).

In the fourth edition of the *APA Publication Manual*, that *p*-values are not acceptable indices of effect was emphasized for the first time, and researchers are "[therefore] encouraged to provide effect-size information" (APA, 1994, p. 18). The Task Force on Statistical Inference was formed by APA to examine prevailing statistical practices, including statistical significance testing. The new recommendations emphasized that "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). The fifth edition of the *APA Publication Manual* (APA, 2001) further recommended reporting effect size measures along with statistical significance testing "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" (p. 26). In June 2006, the AERA published the standards for reporting on empirical social science research, recommending authors to include an index of effect size, standard error and/or confidence interval, and qualitative interpretation of the effect size for each statistical result that is critical to the logic of the design and analysis (AERA, 2006). In the sixth edition of the *APA Publication Manual* (APA, 2010), reporting effect size and confidence intervals along with NHST was further emphasized.

Purpose of the Study

A number of studies were conducted to investigate effect size reporting practice in academic journals from 1967 to 2004 (Alhija & Levy, 2007; Dar et al., 1994; Dunleavy, Barr, Glenn, & Miller, 2006; Hutchins & Henson, 2002; Ives, 2003; Keselman et al., 1998; Kirk, 1996; Lance & Vacha-Haase, 1998; McMillan, Lawson, Lewis, & Synder, 2002; Meline & Schmitt, 1997; Meline & Wang, 2004; Ottenbacher & Barrett, 1989; Paul & Plucker, 2003; Plucker, 1997; Snyder & Thompson, 1998; Thompson, 1999b; Thompson & Snyder, 1997; Vacha-Haase & Ness, 1999; Vacha-Haase & Nilsson, 1998; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000; Ward, 2002). It was consistently found that effect size reporting practice varies across different analyses employed. Specifically, multivariate analyses are more likely to contain effect size than univariate analyses (Alhija & Levy, 2007; Hutchins & Henson, 2002; Ives, 2003; Paul & Plucker, 2003). However, inconsistent conclusions were also identified. For example, Dunleavy et al. (2006) found that variance-accounted-for effect sizes were typically omitted, but Kirk (1996) and McMillan et al. (2002) concluded that R^2 was the most often used effect size measure.

Most of the previous studies emphasized on effect size reporting practices rather than interpreting practices. It has been argued that it is insufficient to simply report effect size statistics, and the researchers need to interpret them as well (Keselman et al., 1998; Thompson, 1996); therefore, the purpose of the present study was to update the investigation into year 2007 to further examine the trend across time with a focus on interpretation of effect size. Specifically, four research questions were addressed in the present study. First, what was the frequency of reporting effect size, and what types of effect size measures were more frequently reported than others? Second, what was the frequency of interpreting effect size, and what types of effect size measures were more frequently interpreted than others? Third, was there any discrepancy between statistical significance and practical significance of the results? If yes, did the authors address the discrepancy? Last, was there any difference in effect size reporting and interpreting practices between different statistical methods, journal sponsors, and publication years?

Method

Data Source

Purposeful sampling strategy was used to select journals to be included in the present study. Purposeful sampling selects information-rich cases for in-depth study; sample size and specific cases depend on the purpose of the study (Patton, 1990, 1999). The main purpose of the present study was to investigate the effect size reporting and interpreting practices in education and psychology areas that are heavily influenced by two professional associations, APA and AERA. As discussed earlier, both APA and AERA have played a significant role in urging the reporting and interpretation of effect sizes; thus, it is important to review the practices of journals sponsored by these two organizations. Journals not sponsored by any academic organizations, labeled as independent journals in the present study, were also included to serve as a comparison for APA and AERA journals. See Table 1 for a complete list of reviewed journals.

Out of the six journals sponsored by AERA, two journals—*Educational Evaluation and Policy Analysis* and *American Educational Research Journal*—were selected because of their high proportion of quantitative empirical research. The official website of APA offers subject guides for all the journals, and six journals by subject "Cognitive/Learning/Education" were selected (http://www.apa.org/journals/by_subject.html). Additional six journals that are not affiliated with any academic associations were taken from the original journal list of Alhija and Levy's (2007) study; those journals were frequently reviewed in the previous studies, thus it allows comparing the trends across different publication years. All the articles published in those journals from 2005 to 2007 were reviewed with the exclusion of articles doing secondary analyses, book reviews, editorials, and journal announcements to ensure the validity of the review for the present sample. Among all of the reviewed quantitative studies, those utilizing NHST were included in this present study.

Review and Coding Procedures

A 17-item checklist adapted from Alhija and Levy (2007) was used as the instrument for this present study (see the Appendix). Major statistical method used in each article was reviewed as per the checklist. *Major statistical method* was defined as the method that is directly used to address the research questions. After all the eligible articles were reviewed, the raw data were coded into different categories for each variable by the first author. On the basis of their sponsors, journals were coded into three categories: AERA journals, APA journals, and independent journals. On the basis of their natures and model complexity, all the NHST methods

Table 1
Reviewed Journals and Their Types

No.	Journal name	Type
1	*Educational Evaluation and Policy Analysis*	AERA journal
2	*American Educational Research Journal*	AERA journal
3	*Journal of Educational Psychology*	APA journal
4	*Journal of Experimental Psychology: Applied*	APA journal
5	*Journal of Experimental Psychology: Human Perception and Performance*	APA journal
6	*Journal of Experimental Psychology: Learning, Memory, and Cognition*	APA journal
7	*School Psychology Quarterly*	APA journal
8	*Dreaming*	APA journal
9	*Early Childhood Research Quarterly*	Independent journal
10	*Journal of Experimental Education*	Independent journal
11	*Journal of Learning Disabilities*	Independent journal
12	*Learning Disabilities Research & Practice*	Independent journal
13	*Journal of Special Education*	Independent journal
14	*Infant and Child Development*	Independent journal

Note. AERA = American Educational Research Association; APA = American Psychological Association.

were coded into three categories: (a) *simple tests*, which include, but are not limited to, *t*-tests, correlation, chi-square, and Mann–Whitney *U* test; (b) *general linear models*, which include analysis of variance (ANOVA), analysis of covariance, multivariate analysis of variance, multivariate analysis of covariance, and regression; and (c) *complex models*, which include structural equation models and hierarchical linear models. On the basis of the classification of effect size measures by Kirk (1996), the effect size measures were coded into three categories: (a) measures of mean differences that include, but are not limited to, Cohen's *d*, *f*, *h*; (b) measures of strength of association that include, but are not limit to, *r*, *R*-squared (R^2), and eta-squared (η^2); and (c) others that include odds ratio, Cohen's *U1*, *U2*, *U3*; or a mixed use of measures from the previous two categories. The practices of each article were coded into dichotomous categories: (a) effect size reported versus not reported, (b) effect size interpreted versus not interpreted, (c) discrepancy between *p*-value and effect size existed versus not existed, and (d) discrepancy between *p*-value and effect size addressed versus not addressed.

It is worthy to note that interrater reliability, which is crucial to studies involving subjective judgment in scoring/rating/coding process, was not reported in this study because the present study involved very little subjective judgment (e.g., reported vs. not reported, interpreted vs. not interpreted), and most of the coded variables are dichotomous.

Data Analysis

All the coded variables were nominal scales and did not produce numerical values that can be used to calculate means and variances; therefore, nonparametric test based on chi-square statistic was applied to analyze the data. All of the tests were conducted at $\alpha = .05$. Cramer's *V* was also reported and interpreted as the effect size measure. Cramer's *V* is used to measure the correlation for data consisting of two categorical variables that have more than two levels, and it ranges from 0 to 1 (Gravetter & Wallnau, 2007; Kline, 2004). As discussed earlier in the present study, the interpretation of effect size is context-dependent, thus Cohen's guidelines are generally not helpful. However, the present study can be considered as an area where no external criteria are established to interpret the importance of the effect size; therefore, Cohen's benchmarks do help the interpretation of Cramer's *V*. As Cohen (1988) suggested, for chi-square tests with degrees of freedom equal to 2, a value of Cramer's *V* within the range of .07–.21 indicates a small effect, a value within the range of .21–.35 indicates a medium effect, and a value larger than .35 indicates a large effect.

Besides Cohen's approach, in the present study we took one step further to explore an alternative way to make meaningful interpretation of Cramer's *V* on the basis of careful examination of the properties of Cramer's *V* as an effect size measure and the nature of the data. First of all, Cramer's *V* is a symmetric and margin-bound measure (Kline, 2004). Changing the definitions of the levels of a categorical variable can change its value (Seaman, 2001). Cramer's *V* reaches its maximum value of 1 when the marginal proportions for rows and columns are equal; as the row and column marginal proportions diverge, the minimum value approaches zero (Kline, 2004). This implies that a value of *V* will change when the cell frequencies in any row or column are multiplied by an arbitrary constant. As shown in Fleiss (1994), a value of *V* is heavily influenced by the ratio of row or column counts; thus, it is uninformative when sampling is not random (Kline, 2004). In addition, although Cramer's *V* is a generalization of φ, the correlation coefficient for dichotomous variables, Cramer's *V* is technically not a correlation coefficient, and its square cannot be interpreted as the proportion of variance explained (Kline, 2004). These properties of Cramer's *V* imply that it is more reasonable to compare proportions than to focus solely on values of Cramer's *V*. If a *V* is used as a measure of a relationship, it should be interpreted as a secondary index in conjunction with a discussion of proportional differences (Seaman, 2001). Therefore, the explicit interpretation of effect size in the present study primarily relies on proportions to locate the differences in the contingency table. Interested researchers can refer to Cohen's guidelines to interpret the importance of the reported effect size values or to make comparison with future research.

Results and Discussion

Descriptive Statistics

In total, 1,581 empirical articles were published in 189 issues of the 14 journals from 2005 to 2007, and 79% of them ($N = 1,243$) were identified as eligible for the present study. The high percentage shows that quantitative research using NHST dominates educational and psychological research. All of the articles were grouped on the basis of type of the journal, the main NHST method used, and year published. See Table 2 for the number of articles in each category. Of the articles, 69% ($n = 863$) were published in APA journals, whereas only 6% ($n = 69$) were published in AERA journals. Furthermore, 76% ($n = 938$) of the articles used general linear models as the main NHST methods. The numbers of articles across the 3 years do not vary very much; the percentages are 35%, 34%, and 31%, respectively.

Reporting Effect Size

Table 3 summarizes the effect size reporting practices for each category. Of the 1,243 articles, 49% ($n = 610$) reported effect size. There exist statistically significant differences between the three types of journals, $\chi^2(2, N = 1,243) = 84.70$, $p = .00$, Cramer's $V = .26$. Because the groups are arbitrarily created and because the sample is not random, Cramer's V may not directly indicate the strength of relationship between journal types or whether effect size is reported. As suggested by Seaman (2001), the interpretation of this result should be complemented by the differences between proportions. Compared with the other two types of journals, AERA journals performed the best, having the highest effect size reporting rate of 73% and, therefore, the largest proportional difference of 45%. The proportional differences for APA journals and independent journals are −19% and 36%, respectively.

There is no statistical difference in effect size reporting within main NHST method type, $\chi^2(2, N = 1,243) = 4.95$, $p = .08$, Cramer's $V = .06$. As addressed in the Method section, different statistical methods were categorized into three groups on the basis of model complexity; if a different characteristic of statistical models was used to create the levels of NHST method type, Cramer's V could be different from .06 because of the nature of Cramer's V mentioned earlier. Therefore, the result in the present study can be also interpreted on the basis of proportional differences. Of the articles using complex models, 60% reported effect size, followed by 49% for general linear models, and 47% for simple tests; the proportional differences are −7% for simple tests, −3% for general linear models, and 19% for complex models. The proportional differences may indicate that the use of effect size may be related to researchers' amount of knowledge. It is likely that researchers using complex models, such as hierarchical linear models and structural equation models, have more advanced knowledge of statistics and, therefore, are more likely to report effect size; on the other hand, researchers using simple tests, especially the not so popular methods, may not know which effect size measure to use or may ignore the importance of reporting effect size.

There is no statistical difference within publication year, $\chi^2(2, N = 1,243) = 5.66$, $p = .06$, Cramer's $V = .07$. Though the proportional differences between reporting and not reporting are pretty small (−10% for 2005, −2% for 2006, and 7% for 2007), the strictly increasing proportional differences do show a clear and positive trend that reporting effect size is steadily increasing across the 3 years.

Of the 610 articles that reported effect size, the most frequently reported type of effect size measures is measure of strength of relations. See Table 4 for details. This result is consistent with the findings in the previous studies (e.g., Alhija & Levy, 2007; Hutchins & Henson, 2002; Kirk, 1996; McMillan et al., 2002), though Dunleavy et al. (2006) found that variance-account-for statistics were typically omitted. The popularity of this type of effect size measure can be explained by the fact that 76% of the 1,243 articles ($n = 938$) used general linear models as the main NHST method, and 75% of the 610 articles that reported effect size ($n = 455$) used general linear models (cf. Table 2).

Table 2
Number of Articles in Each Category

Category	*n*	%
Journal type		
AERA journals	69	6
APA journals	863	69
Independent journals	311	25
Main NHST method		
Simple tests	204	16
General linear models	938	76
Complex models	101	8
Year published		
2005	438	35
2006	422	34
2007	383	31
Total	1,243	

Note. AERA = American Educational Research Association; APA = American Psychological Association; NHST = null hypothesis significance testing.

Interpreting Effect Size

As discussed earlier in this article, applying Cohen's rule of thumb by indicating whether effect size is small, medium, or large or using equivalent words is the basic and most popular way to interpret effect size (Thompson, 2008). Some articles further provided definition of effect size measure, justification of using this measure, and how to understand the effect size value in the context of the research question. For example, in May and Supovitz's (2006) study, the definition of standardized effect size was provided; the choice of this measure was justified; the cutoff values of small, medium, and large effect were provided; the difference between this standardized effect size and Cohen's d was explained; and how to understand the effect in the context of the research question was discussed. This is a good example of interpreting effect size. Table 5 indicates that only a small proportion of articles provide definition (12%) and justification (4%) of choice of the effect size measures. However, it is reasonable for the authors to assume that the readers are aware of the definitions of the effect size measures, except for some uncommonly used measures. Therefore, in the present study, as long as Cohen's guidelines were applied in an article, it is coded as effect size interpreted. Although

Table 3
Number of Articles That Reported Effect Size and Chi-Square Test Results

	Reported		Not reported						
Category	*n*	%	*n*	%	Total	χ^2	*df*	*p*	Cramer's *V*
Journal type									
AERA journals	50	73	19	27	69	84.7	2	<.001	.26
APA journals	349	40	514	60	863				
Independent journals	211	68	100	32	311				
Main NHST method									
Simple tests	95	47	109	53	204	4.95	2	.08	.06
General linear models	455	49	483	51	938				
Complex models	60	59	41	41	101				
Year published									
2005	198	45	240	55	438	5.66	2	.06	.07
2006	207	49	215	51	422				
2007	205	53	178	47	383				
Total	610	49	633	51	1,243				

Note. AERA = American Educational Research Association; APA = American Psychological Association; NHST = null hypothesis significance testing.

interpretation using Cohen's guidelines may be problematic, the positive side is that it does indicate some authors' awareness of the necessity of interpreting effect size.

Table 6 summarizes the number of articles that interpreted effect size in each category and the chi-square test results. Of the 610 articles that reported effect size, 57% ($n = 346$) contained interpretation. There exist statistically significant differences within journal type, $\chi^2(2, N = 610) = 9.90$, $p = .01$, Cramer's $V = .13$. Independent journals had the highest rate of interpreting effect size, which is 65%, followed by AERA journals (62.0%) and APA journals (51%). The overall rate is 57%, that is, 346 of the 610 articles that report effect size also provide interpretation. There exist statistically significant differences within NHST method type, $\chi^2(2, N = 610) = 7.52$, $p = .02$, Cramer's $V = .11$. Similar to the finding in effect size reporting, articles that employed complex models are more likely to interpret effect size than others, as indicated by the proportional difference of 47% between interpreting versus not interpreting. Within the 3 years, though the chi-square test is statistically nonsignificant, $\chi^2(2, N = 610) = 0.43$, $p = .81$, Cramer's $V = .03$, the proportions of effect size interpreting are quite stable across years, 59% for Year 2005, 56% for Year 2006, and 56% for Year 2007.

Similar to the findings in effect size reporting, the most frequently interpreted effect size measure type is measure of strength of relations, which account for 63% of the 346 interpreted effect size measures ($n = 217$). This is consistent with the fact that general linear models are the most popular methods among the reviewed articles. This result is partially consistent with the findings by Alhija and Levy (2007) that effect size was more frequently interpreted in *t*-test and regression.

Table 4
Frequency of Effect Size Reporting for Different Effect Size Measures

Type of measures	*f*	%
Measure of mean differences	156	26
Measure of strength of relations	383	62
Others	71	12
Total	610	

Discrepancy Between *p*-Value and Effect Size

The present review covers diverse articles in the literature, and established criteria for practical significance may not be available for each area; thus, it is extremely difficult to judge how large an effect is truly meaningful and significant for each article in the review process. Though it is problematic to blindly apply Cohen's benchmarks to judge whether the effect size is small, medium, or large (Glass et al., 1981; Thompson, 1999a, 2008), some scholars argue that Cohen's benchmarks are reasonably accurate (Glass, 1979; Olejnik, 1984). Therefore, in the present study we still use Cohen's benchmarks and consider medium and large effect size as practically significant. It is common that a study has multiple outcomes measures and reports multiple effect sizes; it will be classified as statistically significant as long as at least one effect size value reaches medium size. An article is classified as discrepant when medium or large effect is found for a statistically nonsignificant test or a small effect is found for a statistically significant test. This criterion resulted in 69 out of 610 articles (11%)

Table 5
Frequency of Definition of Effect Size and Justification of Effect Size Choice

	Yes		No		Total	
Variable	*f*	%	*f*	%	*f*	%
Whether the definition of the effect size measure is stated	75	12	535	88	610	100
Whether the choice of the effect size measure is justified	24	4	586	96	610	100

Table 6
Number of Articles That Interpreted Effect Size

	Interpreted		Not interpreted						
Category	*n*	%	*n*	%	Total	χ^2	*df*	*p*	Cramer's *V*
Journal type									
AERA journals	31	62	19	38	50	9.90	2	.01	.13
APA journals	179	51	170	49	349				
Independent journals	136	64	75	36	211				
Main NHST method									
Simple tests	53	56	42	44	95	7.52	2	.02	.11
General linear models	249	55	206	45	455				
Complex models	44	73	16	27	60				
Year published									
2005	116	59	82	41	198	0.43	2	.81	.03
2006	115	56	92	44	207				
2007	115	56	90	44	205				
Total	346	57	264	43	610				

Note. AERA = American Educational Research Association; APA = American Psychological Association; NHST = null hypothesis significance testing.

having a discrepancy between *p*-value and effect size. For example, in Experiment 1 of Vachon, Tremblay, and Jones's (2007) study, the priming effects of Visual Target 2 were compared in two treatment conditions with a switch of location with $n = 15$ participants using a $2 \times 2 \times 3$ repeated measures ANOVA. Two statistically nonsignificant interaction effects were found, the interaction between task switching and target relation, $F(1, 14) = 2.23$, $p = .16$, $d = 0.80$, and the interaction between target relation and lag, $F(2, 28) = 2.27$, $p = .122$, $d = 0.81$. Both *d*s are considered to be a large effect by Cohen (1988).

Three chi-square tests were conducted to investigate the differences of discrepant findings across three types of journals, $\chi^2(2, N = 610) = 1.50$, $p = .47$, Cramer's $V = .05$; three types of NHST methods, $\chi^2(2, N = 610) = 4.06$, $p = .13$, Cramer's $V = .08$; and 3 years, $\chi^2(2, N = 610) = 2.09$, $p = .35$, Cramer's $V = .06$. All of the results are not statistically significant. See more details in Table 7. Because of the loose criterion used to identify discrepant studies in a nonrandom sample, the data reported in Table 7 may not reflect the true situation in education and psychology areas.

Whether the Discrepancy Was Addressed by the Author

For those articles that are identified as having a discrepancy between *p*- value and effect size, if the authors discussed the possible reasons of the discrepancy, the article was classified as "discrepancy addressed." For example, in Simard and Nielsen's (2005) study about dreaming, an analysis of covariance test did not produce a statistically significant result as indicated by $F(2, 40) = 2.42$, $p = .10$; however, the effect size was .46, which was interpreted as a large effect by the authors in the Discussion section. In their Discussion section, the authors explained that the absence of a robust difference was probably due to the small sample size. However, the authors did not explain how the effect

Table 7
Discrepancy Between p-Value and Effect Size

	Discrepancy		No discrepancy						
Category	*n*	%	*n*	%	Total	χ^2	*df*	*p*	Cramer's *V*
Journal type									
AERA journals	8	16	42	84	50	1.50	2	.47	.05
APA journals	36	10	313	90	349				
Independent journals	25	12	186	88	211				
Main NHST method									
Simple tests	16	17	79	83	95	4.06	2	.13	.08
General linear models	45	10	410	90	455				
Complex models	8	13	52	87	60				
Year published									
2005	26	13	172	87	198	2.08	2	.35	.06
2006	25	12	182	88	207				
2007	18	9	187	91	205				
Total	69	11	541	89	610				

Note. AERA = American Educational Research Association; APA = American Psychological Association; NHST = null hypothesis significance testing.

size was computed or what criteria were used to interpret it as a large effect. It should be noted that in some areas, some significantly large treatment effects may be judged as small on the basis of Cohen's benchmarks; therefore, it is crucial for content experts to set up criteria in their fields to decide how large is large, instead of applying Cohen's guidelines blindly.

Of the 69 articles that have discrepant results on the basis of *p*-value and effect size, 30% of them ($n = 21$) were addressed by the authors. Three chi-square tests were conducted to investigate the differences between the three types of journals, $\chi^2(2, N = 69) = 6.01$, $p = .05$, Cramer's $V = .30$; NHST methods, $\chi^2(2, N = 69) = 0.13$, $p = .94$, Cramer's $V = .04$; and 3 years, $\chi^2(2, N = 69) = 2.11$, $p = .47$, Cramer's $V = .18$. See more details in Table 8. Overall, less than one third of articles that found discrepant results addressed the possible reasons. This result suggests that possible threats to the study validity went unnoticed in those studies. Failure to address the discrepancy between *p*-value and effect size may mislead the consumers of the studies, especially researchers who are planning a meta-analysis study.

Conclusion and Limitation

Reporting and interpreting effect size enables the consumers of the studies to have a clear understanding of the size and the meaning of the effect. As a metric-free measure of the size of mean differences or strength of relations, effect size may be used to compare the results of different studies with one another. The previous studies found proportions of effect size reporting ranging from 1% (Meline & Schmitt, 1997) to 87% (Thompson, 1999b), and it was 49% in the present study. The proportions of effect size interpreting was about 40% in Alhija and Levy's (2007) study, 50% in Meline and Wang's (2004) study, 88% in Hutchins and Henson's (2002) study, and 57% in the present study. Because the reviewed journals and review criteria were different among the studies—especially that some previous studies used very small sample sizes because of limited number of journals selected and restricted time span (e.g., Hutchins & Henson, 2002, used a sample size of 14 articles published in 2000, and Thompson & Snyder, 1997, used a sample size of 22 articles published in 1994–1996 in the *Journal of Experimental Education*)—the variations in proportions are reasonably expected. As an attempt to replicate and update the findings in the previous studies, in the present study we reviewed the most recent practices of effect size reporting and interpreting in education and psychology areas; an overall rate of 49% for effect size reporting and 57% for effect size interpreting is still far from satisfactory. Though a positive trend was revealed in effect size reporting, it suggests that sufficient emphasis has not yet been put upon reporting and interpreting effect size.

The present study shows that effect size reporting practice differs between journal types but does not statistically differ between different types of NHST methods, which seems contradictory to the previous studies (Alhija & Levy, 2007; Dunleavy et al., 2006; Hutchins & Henson, 2002; Ives, 2003; Paul & Plucker, 2003). As far as effect size interpreting practice is concerned, it statistically differs in both journal types and types of NHST methods. As far as the frequency of discrepancy between *p*-value and effect size and whether authors address the discrepancy are concerned, statistically nonsignificant results show that they do not differ between types of journals or types of NHST methods. It is reasonable to assume that discrepancy occurs somewhat at random; however, the overall rate of 30% for addressing the discrepancy was low and suggests that researchers should pay attention to this matter when analyzing the data and writing the report. None of the four tests about the time effect were statistically significant, but a noticeably increasing trend was found in effect size reporting practice. The proportion of articles that interpret effect size was relatively stable across 3 years but was still far from sufficient.

The present study also shows that measures of strength of relations are the more likely to be reported and interpreted than the other measures. As Alhija and Levy (2007) proposed that this may be due to the fact that those measures are usually produced automatically by software and handled more often in statistical education coursework as well, it cannot be totally interpreted as high awareness of the importance of reporting and interpreting effect size measures. The popularity of measures of strength of

Table 8
Whether the Authors Address the Discrepancy Between p-Value and Effect Size

	Yes		No						
Category	*n*	%	*n*	%	Total	χ^2	*df*	*p*	Cramer's *V*
Journal type									
AERA journals	1	12	7	88	8	6.01	2	.05	.30
APA journals	8	22	28	78	36				
Independent journals	12	48	13	52	25				
Main NHST method									
Simple tests	5	31	11	69	16	0.13	2	.94	.04
General linear models	14	31	31	69	45				
Complex models	2	25	6	75	8				
Year published									
2005	9	35	17	65	26	2.11	2	.35	.18
2006	5	20	20	80	25				
2007	7	39	11	61	18				
Total	21	30	48	70	69				

Note. AERA = American Educational Research Association; APA = American Psychological Association; NHST = null hypothesis significance testing.

relations is consistent with the fact that general linear models are the most frequently used NHST methods in the present study. In the statistical education coursework, measures of strength of relations, such as R^2 in regression family, are probably given more attention than other measures, thus researchers are more comfortable to use them.

The result that 11% of the 610 articles that reported effect sizes had discrepant results on the basis of *p*-value and effect size measures also needs to be interpreted with caution because of the loose classification criterion employed by the present study. However, among the 69 articles that had discrepant results, only 30% of them ($n = 21$) explained the possible reasons for the discrepancy. This low percentage is consistent with the findings in Alhija and Levy's (2007) study. Many researchers tend to ignore the meaning and importance of effect size measures in the context of their research or the quality of their results. In a majority of the 21 articles, the discrepancy was casually addressed by saying that the *p*-value is significant but effect size is very small, and therefore, the results should be interpreted with caution. Only a few studies explained why the discrepancy occurred (e.g., Simard & Nielsen, 2005). Discrepancy between the *p*-value and effect size measures can be produced by several reasons, such as inadequate sample size and violation of the assumptions of the NHST methods; therefore, this pertains to the importance of conducting prior power analysis, checking research design, and the quality of the data.

The limitation of the present study is the use of purposeful sampling. Though this strategy best serves the purpose of in-depth investigation of effect size reporting and interpreting practices in education and psychology, the 14 journals are not a random sample of the entire population and, therefore, limit the generalizability of the results. Because of the violation of random sampling assumption, Cramer's *V* estimated in this study as the effect size for chi-square test is heavily influenced by the relative cell frequencies. Thus, the explicit interpretation of the results reported in the present study primarily relied on descriptive statistics, and Cramer's *V* was considered as a secondary effect size measure. Cramer's *V* was explicitly reported for all of the chi-square tests, and Cohen's cutoff values to interpret effect size were provided. Thus, the explicit interpretation of proportional differences does not prevent interested researchers to interpret the importance of the effects or to compare *V* across studies but instead makes the interpretation richer and more meaningful.

An Example of Reporting and Interpreting Effect Size

Among all of the articles reviewed in the present study, Jitendra et al. (2007) did a good job in reporting and interpreting effect size. The purpose of their study was to investigate the differential effects of a single strategy (schema-based instruction [SBI]) versus multiple strategies (general strategy instruction [GSI]) in promoting mathematical problem solving. Because of the experimental nature of their study, they reported effect size measures of mean differences. There are three noteworthy strengths of reporting and interpreting effect size in their study. First, how effect sizes were computed was clearly explained with references. Second, each reported effect size value was interpreted by indicating whether it is small, medium, or large. Third, the effect sizes in their study were explicitly and directly compared with effect sizes found in previous studies. Below is a summary of how they reported and interpreted one of the effect sizes in their study.

For posttest comparison, regressed adjusted mean difference divided by square root of mean square error (Glass et al., 1981) was reported as an effect size measure. After controlling for the SAT-9 scores, the SBI group ($n = 45$, $M = 20.92$, $SD = 7.05$) performed better than the GSI group ($n = 43$, $M = 18.59$, $SD = 7.36$) in the posttest on word problem solving, $F(1, 84) = 5.96$, $p < .05$, *effect size* = 0.52, which is a medium effect. This medium effect size corroborates previous studies regarding the effectiveness of SBI in solving arithmetic word problems, but it is considerably smaller than the large effects (range = 1.55–3.72) found in previous studies (e.g., Fuchs, Fuchs, Finelli, Courey, & Hamlett, 2004; Fuchs, Fuchs, Prentice, et al., 2004). The inconsistent findings across studies could be explained by the fact that Jitendra et al. (2007) used a more robust comparison condition than the control condition in the previous study.

It would be better if the authors could explain the substantive significance of effect size 0.52, beyond Cohen's guidelines, in the context of teaching mathematical problem solving in third grade classrooms. Mathematically, an effect size of 0.52 indicates that the SBI group outperformed the GSI group by 0.52 standard deviations, but what does this number mean for researchers and third grade mathematics teachers? How impressive the effect size is in the context of the research? These are the questions that the audiences are interested in and that the content experts should answer.

Practical Guidelines for Researchers

Researchers' resistance to reporting effect size may be partially explained by some combination of confusion and desperation about NHST and effect size (Thompson, 1999d). One of the sources of researchers' confusion may come from textbooks. Two review studies on statistics books were conducted in 2002. R. M. Capraro and Capraro (2002) reviewed textbooks published from 1995 to 2002 on treatments of effect size and statistical significance tests. Of the textbooks examined, all textbooks ($N = 89$) included the topic of statistical significance testing (2,248 pages), whereas only a little more than two thirds of the textbooks ($n = 60$) included information about effect sizes (789 pages). Obviously insufficient attention was given to effect size compared with NHST. Curtis and Araki (2002) conducted another review of 22 statistics textbooks in education and psychology areas to examine the ways in which authors addressed the issue of effect size and the practical significance of research results. The identified problems with the way to present effect size statistics include the failure to distinguish between effect size parameters and statistics, the use of conceptually uninformative formulas, and the lack of agreement on how to calculate specific effect size statistics. Problems with the ways the authors discussed the interpretation of effect size statistics were also identified. Statistics textbooks are the tools of researchers, students, and future researchers; therefore, those problems with effect size in textbooks probably affect or will affect their practice in research. To alleviate researchers' confusion about effect size, exemplary studies on how to understand different types of effect size measures would significantly contribute to the literature.

In addition to the tutorials that can be found in Cromwell (2001), Glass (1976), Hojat and Xu (2004), Kirk (1996), Lakshmi (2000), Mahadevan (2000), Robey (2004), Smithson (2001), Snyder and Lawson (1993), and Volker (2006), researchers are strongly recommended to refer to the following nine annotated studies to select, report, and interpret effect size measures and confidence intervals appropriately.

1. Thompson (2002b, 2007, 2008): The importance and utilization of effect size, confidence intervals, and confidence intervals for effect size are thoroughly addressed using formulas, empirical data, and graphic demonstration.

2. Kampenes, Dybå, Hannay, and Sjøberg (2007): This is good summary of how to understand and use standardized effect size, unstandardized effect size, and nonparametric effect size; it is primarily written for software engineers but motivated by the literature in education and psychology.

3. Nakagawa and Cuthill (2007): How to calculate effect size and construct confidence intervals on the basis of statistical models is addressed; how to deal with bias of effect size and the violation of assumptions are explicitly discussed.

4. Henson (2006): Explicit examples are provided for reporting and interpreting effect size; meta-analytical thinking or assessment for effect size is strongly recommended.

5. Kline (2004): This is an excellent book on how to reform data analysis methods in behavioral research, including fundamental concepts and problems with NHST, estimating effect size in comparative studies, and alternatives to statistical tests.

6. Trusty, Thompson, and Petrocelli (2004): This is a very practical guide for use of effect size on the basis of different statistical models with a brief instruction for SPSS.

7. Vacha-Haase and Thompson (2004): This is a helpful tutorial on how to estimate and interpret effect size with detailed strategies for obtaining effect size for some statistical models in SPSS.

8. Huberty and Lowman (2000): This is an excellent article that explicitly interprets effect size for mean differences as group overlaps.

9. Sim and Reid (1999): This is all about confidence intervals that quantitative researchers should know.

To make the presentation of effect size and other evidence of study validity and usefulness clear, sufficient, and conform to the reporting standards set up by APA and AERA, we provide the following guideline. This guideline considers the whole process of an empirical study and how to incorporate the use of effect size into the process.

Step 1: In the Research Design Phase, Estimate a Sample Size That Is Sufficient to Detect a Meaningful Effect Size Proposed by Previous Studies

The purpose of effect size estimation is to make sure that the low statistical power is not a threat to statistical nonsignificance and to avoid the discrepancy between *p*-value and effect size. This step is known as prior power analysis in the literature. There is another type of power analysis named post hoc power analysis using obtained effect size to estimate the achieved power of the study. This type of power analysis assumes that the observed effect size is exactly equal to population effect size then estimates the power of the test using observed effect size, alpha level, and sample size. It is recommended as supplementary information to the validity of a test, especially when the result is nonsignificant. However, high post hoc power does not confirm that the theory that guides the test is correct; low post hoc power may be due to small effect size instead of small sample size. Moreover, nonsignificant tests always have low statistical power. Therefore, a post hoc power analysis is "more like an autopsy than a diagnostic procedure. That is, it is better to think about power before it is too late" (Kline, 2004, p. 43). For more discussion about the misuse and misinterpretation of post hoc power analysis, see Colegravea and Ruxton (2003); Hoenig and Heisey (2001); Sun, Pan, and Wang (2009); Thomas (1997); and Yuan and Maxwell (2005).

For a mature field or well-studied research question, the effect size used to estimate the sample size should come from prior empirical studies or meta-analysis. This is exactly where meta-analytical assessment of effect size starts. For a premature field or research question, Cohen's three-number guideline may still be helpful to estimate the sample size because it is reasonably accurate (Glass, 1979; Olejnik, 1984).

Step 2: Check the Assumptions and Clean Data Before Data Analysis

Assumptions are very critical for statistical model selection and application. Both statistical methods and effect size measures have different tolerances to assumption violations. Thus, researchers should check the assumptions before data analysis, and severe violation of assumption should be remedied and discussed in the manuscript to justify the choice of statistical methods and effect size measures. Also, the data set should be cleaned before data analysis to make sure that missing data and possible outliers are appropriately handled.

Step 3: Report Descriptive Statistics Before Inferential Statistics

For parametric statistical methods, sample size, or group sizes, means and standard deviations should be reported. Means and standard deviations are sufficient statistics that contains enough information about the sample, and their richness cannot be replaced by inferential statistics, such as t or F. For instance, if ANOVA is used to analyze data collected from a three-group experiment, omnibus F value cannot tell researchers which group is different from others; by looking at the means and standard deviations, researchers can reasonably determine which treatment makes a difference before conducting post hoc contrast. For nonparametric statistics, counts and percentages should be reported. Reporting detailed descriptive statistics not only makes the study clear and easy to read but it also facilitates meta-analysis in the future. When multiple outcomes are investigated in a study, the authors may consider using tables to present all descriptive and inferential statistics. Clearness and richness of the statistics should not be sacrificed.

Step 4: Report Exact *p*-Value and Degrees of Freedom for Inferential Statistics

The dichotomous decision pattern of NHST is strictly reinforced by the current use of "$p < .05$" or "$p < .01$." Though *p*-value cannot tell the practical significance of the effect, there exists a big

difference between $p = .049$ and $p = .011$ in terms of the probability of the sample statistics in its sampling distribution. The sample associated with $p = .011$ is much less possible in the sampling distribution than the sample associated with $p = .049$. However, both p-values can be rounded up to $p < .05$, and the richness of information is lost. Therefore, the current pattern of p-value reporting should be replaced by exact p-value. In addition, degrees of freedom are directly related to the decision of the critical region of the test and, thus, should be explicitly reported.

Step 5: Report Confidence Intervals for Parameter Estimates and Effect Size Right After the *p*-Value, and the Name of the Effect Size Should Be Explicitly Stated

Whether significant, p-values should always be followed by confidence intervals of the parameter estimates and the name and value of the effect size measures. The advantages of reporting confidence intervals for parameter estimates have been discussed earlier in this article. The purpose of reporting effect size for both significant and nonsignificant p-values is to further check whether discrepancy between p-value and effect size exists. Multiple effect size measures could be used for the same test, thus researchers should explicitly tell readers the name of that measure instead of using the broad name "effect size." In the present study, about 7% ($n = 42$) of the 610 articles that reported effect size used "effect size" instead of the name of that measure. This is what researchers should avoid when writing the manuscript. Different effect size measures have different properties, different ranges of values, and different interpretations; without telling readers the name of that measure, researchers cannot make sense of the study. Authors are also recommended to justify their choices of effect size measures, especially when the choice is heavily influenced by the special features of the study or when the measure is uncommon. In the present review, only 4% of the reviewed studies (cf. Table 5) provided justifications to their choice. Justification makes the study rigorous, convincing, and easy to understand and communicate.

Step 6: Effect Size Should Be Explicitly and Directly Interpreted

As discussed earlier in the present study, researchers should meta-analytically interpret the effect size; in other words, researchers should consider within-study interpretation and across-study interpretation of effect size. First, within the study, indicate what the effect size value means for the research topic. It is common that multiple effect sizes are reported in a study, and it may not be feasible to interpret each single effect size value in detail; however, the values that directly speak to the research questions should be interpreted explicitly.

Second, as recommend by Henson (2006), Kline (2004), and Thompson (2008), researchers should explicitly and directly compare the effect size in the study with the prior effect sizes in the related literature whenever possible. This is a cross-study interpretation. Reporting effect size is still not a common practice in many areas, so researchers may find it difficult to find effect sizes reported in previous studies. As long as research designs are clearly communicated and descriptive statistics sufficiently reported, it is still possible to estimate an effect size.

Step 7: Direct the Interpretation of Effect Size to Practice and Future Research

The purpose of measuring effect size is to quantify the actual size of the treatment; as a measure for practical significance, effect size obtained in a study, especially effect size that is consistent across studies, should be directly related to its contribution to educational and psychological practice. Any threat to the validity and effect size of the study should be explicitly discussed so that these threats could be effectively controlled in future research. Content experts are suggested to set up criteria for interpreting effect size on the basis of the comparison of effect sizes across studies and meta-analytical studies to facilitate future research.

References

Alhija, F. N., & Levy, A. (2007, April). *Effect size reporting practices in published articles.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

American Educational Research Association. (2006). Standards on reporting on empirical social science research in AERA publications. *Educational Researcher, 35,* 33–40.

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.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th 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 chi-squared test. *Journal of the American Statistical Association, 33,* 526–536.

Biskin, B. (1998). Comment on significance testing. *Measurement and Evaluation in Counseling and Development, 31,* 58–62.

Capraro, M. M. (2005). An introduction to confidence intervals for both statistical estimates and effect sizes. *Research in the Schools, 12,* 22–32.

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

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

Chow, S. L. (1996). *Statistical significance.* Thousand Oaks, CA: Sage.

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

Colegravea, N., & Ruxton, G. D. (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology, 14,* 446–450.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods, 2,* 161–172.

Cromwell, S. (2001, February). *An introductory summary of various effect size choices.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

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. (2002). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

Curtis, D. A., & Araki, C. J. (2002, April). *Effect size statistics: An analysis of statistics textbooks used in psychology and education.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75–82.

Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. R. (2006). Effect size reporting in applied psychology: How are we doing? *The Industrial-Organizational Psychologist, 43,* 29–37.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94,* 275–283.

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

Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. *American Educational Research Journal, 41,* 419–445.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third grade students with schema-based instruction. *Journal of Educational Psychology, 96,* 635–647.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V. (1979). Policy for the unpredictable (uncertainty research and policy). *Educational Researcher, 8,* 12–14.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gravetter, F. J., & Wallnau, L. B. (2007). *Statistics for the behavior sciences* (7th ed.). Belmont, CA: Wadsworth.

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

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

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

Henson, R. K. (2006). Effect size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist, 34,* 601–629.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. *Journal of Research and Development in Education, 33,* 285–296.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Hojat, M., & Xu, G. (2004). Statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education, 9,* 241–249.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60,* 543–563.

Hutchins, H. M., & Henson, R. K. (2002, February). *In search of OZ: Effect size reporting and interpretation in communication research.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Ives, B. (2003). Effect size use in studies of learning disabilities. *Journal of Learning Disabilities, 36,* 490–504.

Jitendra, A. K., Griffin, C. C., Haria, P., Leh, J., Adams, A., & Kaduvettoor, A. (2007). A comparison of single and multiple strategy instruction on third-grade students' mathematical problem solving. *Journal of Educational Psychology, 99,* 115–127.

Johnson, D. H. (1995). Statistical sirens: The allure of nonparametrics. *Ecology, 76,* 1998–2000.

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology, 49,* 1073–1086.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication, 23,* 251–266.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

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 practice: Some suggestions. *Educational and Psychological Measurement, 61,* 213–218.

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

Lakshmi, M. (2000, January). *The effect size statistics: Overview of various choices.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Lance, T. S., & Vacha-Haase, T. (1998, August). *The counseling psychologist: Trends and usages of statistical significance testing.* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Mahadevan, L. (2000, January). *The effect size statistic: Overview of various choices.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

May, H., & Supovitz, J. A. (2006). Capturing the cumulative effects of school reform: An 11-year study of the impacts of America's choice on student achievement. *Educational Evaluation and Policy Analysis, 28,* 231–257.

McLean, E. J., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5,* 15–22.

McMillan, J. H., Lawson, S., Lewis, K., & Synder, A. (2002, April). *Reporting effect size: The road less traveled.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391–423). Mahwah, NJ: Erlbaum.

Meline, T., & Schmitt, J. F. (1997). Case studies for evaluating statistical significance in group designs. *American Journal of Speech-Language Pathology, 6,* 33–41.

Meline, T., & Wang, B. (2004). Effect-size reporting practices in AJSLP and other ASHA journals, 1999–2003. *American Journal of Speech-Language Pathology, 13,* 202–207.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Review, 82,* 591–605.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. *Journal of Experimental Education, 53,* 40–48.

Ottenbacher, K. J., & Barrett, K. A. (1989). Measures of effect size in the reporting of rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 68,* 52–58.

Patton, M. Q. (1990). *Qualitative evaluation and research methods.* Thousand Oaks, CA: Sage.

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 34,* 1189–1208.

Paul, K. M., & Plucker, J. A. (2003). Two steps forward, one step back: Effect size reporting in gifted education research from 1995–2000. *Roeper Review, 26,* 68–72.

Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review, 2,* 122–126.

Robey, R. R. (2004, November). *Effect sizes in research manuscripts: Selecting, calculating, reporting and interpreting.* Seminar presented before the annual conference of the American Speech-Language-Hearing Association, Philadelphia, PA.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26,* 21–27.

Robinson, D. H., & Wainer, H. (2002). On the past and future of null hypothesis significance testing. *Journal of Wildlife Management, 66,* 263–271.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York, NY: Cambridge University Press.

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

Schmidt, F., & Hunter, J. E. (1995). The impact of data-analysis methods on cumulative research knowledge: Statistical significance testing, confidence intervals, and meta-analysis. *Evaluation and the Health Professions, 18,* 408–427.

Seaman, M. A. (2001). *Categorical data* [Electronic resource]. Retrieved from http://www.ed.sc.edu/seaman/edrm711/questions/categorical.htm

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61,* 350–360.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61,* 293–316.

Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals. *Physical Therapy, 79,* 186–196.

Simard, V., & Nielsen, T. A. (2005). Sleep paralysis–associated sensed presence as a possible manifestation of social anxiety. *Dreaming, 15,* 245–260.

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

Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal. *School Psychology Quarterly, 13,* 335–348.

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.

Sun, S., Pan, W., & Wang, L. (2009, April). *Rethinking observed power: Concept, practice, and implications.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology, 11,* 276–280.

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

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

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

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26,* 29–32.

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

Thompson, B. (1998b). Statistical significance and effect size reporting: Portrait of possible future. *Research in the Schools, 5,* 33–38.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology, 82,* 165–181.

Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65,* 329–337.

Thompson, B. (1999c). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory and Psychology, 9,* 191–196.

Thompson, B. (1999d). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to change practices. *The Journal of Psychology, 133,* 133–140.

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

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 45,* 423–432.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 246–262). Thousand Oaks, CA: Sage.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. *Journal of Experimental Education, 66,* 75–83.

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the *Journal of Counseling & Development*. *Journal of Counseling & Development, 82,* 107–112.

Vacha-Haase, T., & Ness, C. M. (1999). Statistical significance testing as it relates to practice: Use within professional psychology: Research and practice. *Professional Psychology: Research and Practice, 30,* 104–105.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling and Development, 31,* 46–57.

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.

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

Vachon, F., Tremblay, S., & Jones, D. M. (2007). Task-set reconfiguration suspends perceptual processing: Evidence from semantic priming during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 330–347.

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.

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

Ward, R. M. (2002). *Highly significant findings in psychology: A power and effect size survey.* Unpublished doctoral dissertation, University of Rhode Island.

Wilkinson, L., & APA 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.

Yetkiner, Z. E., Capraro, R. M., Zientek, L. R., & Tompson, B. (2008, July). *Effect size and confidence interval reporting practices in mathematics education.* Paper presented at the 11th International Congress on Mathematical Education, Monterrey, Mexico.

Young, M. A. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research, 36,* 644–656.

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141–167.

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel report. *Educational Researcher, 37,* 208–216.

Appendix

Checklist for Coding the Articles

No.	Item	Note/Check
1	Title	_____
2	Year	_____
3	Source	_____
4	Journal sponsor	_____
5	Research questions	_____
6	Major analysis	_____
7	Result (statistically significant/not significant/mixed)	_____
8	Did the author specify "statistically significant" for the result?	□ Yes □ No
9	Are effect sizes reported?	□ Yes □ No
10	Are effect sizes reported also for not significant results?	□ Yes □ No
11	What is the effect size measure?	_____
12	What is the definition of the effect size measure?	_____
13	Is the use of a specific effect size justified?	□ Yes □ No
14	Is the effect size interpreted?	□ Yes □ No
15	Is there a discrepancy between conclusions based on statistical as opposed to practical significance?	□ Yes □ No
16	If such a discrepancy exists, has the author address it?	□ Yes □ No
17	Is practical implication of the study discussed?	□ Yes □ No

Received August 15, 2008
Revision received March 18, 2010
Accepted March 18, 2010 ■