UNIVERSITY OF CINCINNATI

Date: May 20, 2008

I, Shuyan Sun,

hereby submit this work as part of the requirements for the degree of:

Master of Arts

in:

Educational Studies

It is entitled:

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

This work and its defense approved by:

Chair: Wei Pan, Ph.D.

Leigh Wang, Ph.D.

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

A thesis submitted to the

Division of Research and Advanced Studies of the University of Cincinnati

In partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

In Educational Studies Program,

College of Education, Criminal Justice and Human Services

2008

By

Shuyan Sun

Bachelor of Philosophy, Beijing Normal University, 2004

Committee Chair: Wei Pan, Ph.D.

Committee Member: Leigh Wang, Ph.D.

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

Abstract

Reporting effect size to supplement *p*-value in null hypothesis significance testing (NHST) is highly recommended by scholars, journals and academic associations. The current comprehensive review investigated the most recent effect size reporting and interpreting practices of 1,243 studies published in 14 academic journals from 2005 to 2007. Overall, 49.1% of the articles reported effect size and 56.7% of them interpreted effect size. A series of Chi-square tests suggested that (a) effect size reporting and interpreting practices statistically differ between types of journals; (b) only effect size interpreting practice differs between different NHST methods; (c) neither effect size reporting nor interpreting practice differ between years. The importance of reporting and interpreting effect size is also discussed.

Acknowledgement

I would like to express my gratitude to my advisor and chairman, Dr. Wei Pan, for his constant guidance, support, and encouragement throughout my course of study. This thesis would have not been possible without his great work and valuable inputs. I am also sincerely grateful to my committee member Dr. Leigh Wang. Her excellent insight and enthusiasm for research have been very important in my training and development as a quantitative research methodologist; her brilliant comments greatly improved the presentation and content of my thesis.

This thesis is dedicated to my beloved parents and sister in China whom I have been far apart from for two years. Their deep love, high expectation and self-giving support motivate me to study overseas.

Table of Contents

List of Tables

List of Figures

Introduction

Null hypothesis significance testing (NHST) is the traditional and popular approach to make statistical inference about research questions (Anderson, Burnham, & Thompson, 2000). However, the logic and usefulness of NHST have been challenged in the literature (Anderson, Burnham, & Thompson, 2000; Carver, 1978; Cohen, 1990, 1994; Falk & Greenbaum, 1995; Harlow, Mulaik, & Steiger, 1997; Henson & Smith, 2000; Kirk, 1996; Robinson & Wainer, 2002; Schmidt, 1996; Yates, 1951). Effect size measures as a criterion for practical significance has been recommended to supplement NHST to get better statistics and results for a long time (American Educational Research Association, 2006; Anderson, Burnham, & Thompson, 2000; American Psychological Association, 2001; Kirk, 1996; Plucker, 1997; Robinson & Levin 1997; Thompson & Snyder, 1997). The effectiveness of this recommendation is worthy of a methodological review. Thus, the purpose of the present study is to investigate the effect size reporting and interpreting practices in academic journals in education and psychology areas, examine how researchers are doing and further raise the awareness of importance of effect size.

Null Hypothesis Significance Testing

The origin of NHST. The history of NHST can be dated back to 1710 when John Arbuthnot used this procedure to study birth rate. It was popularized in the social sciences by the great efforts of Ronald Fisher, Jerzy Neyman, and Egon Pearson (cf., Thompson, 1996). The present-day NHST is a hybrid of Fisher's significance testing and Neyman and Pearson's hypothesis testing. Most of the ideas underlying NHST, including the theory of point estimation, consistency, efficiency, sufficiency, randomization, and maximum likelihood estimation had been set forth by Fisher in 1925. The ideas of the present-day NHST—Type I and Type II errors

and a predetermined level of significance alpha—were contributed by Neyman and Pearson in 1928 (cf., Kirk, 1996).

NHST frames the research question in terms of two contrasting statistical hypotheses: "The null hypothesis (H_0) 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), while the alternative hypothesis (H_1) states the opposite. These hypotheses correspond to different models in various circumstances. Applying the procedure yields a value of *p*, the theoretical probability that if the samples used had been drawn randomly from the same population that characterizes the null state, the statistical test would have yielded a statistic large or larger than the one obtained. Alpha, a designated specified significance level acts as a decision criterion, and the null hypothesis is rejected only if the *p*-value yielded by the test is not greater than the value of alpha.

Problems with NHST. NHST was considered to be an objective, scientific procedure for knowledge accumulation (Kirk, 1996). However, for almost 80 years, it held a controversial status in social and behavioral research: On 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 challenge to NHST dated back to 1938 when Joseph Berkson published his article to challenge the logic and usefulness of NHST (as cited in Kirk, 1996). Since then, criticisms of NHST have noticeably intensified (Anderson, Burnham, & Thompson, 2000; Carver, 1978; Cohen, 1990, 1994; Falk & Greenbaum, 1995; Harlow, Mulaik, & Steiger, 1997; Henson & Smith, 2000; Katzer & Sodt, 1973; Schmidt, 1996; Yates, 1951). The fundamental problem with the NHST is not that it is wrong, but that it is uninformative in most cases, and of relatively little use in model or variable selection; the interpretation and application

of the results are always problematic issues (Anderson, Burnham, & Thompson, 2000; Ives, 2003).

The problems NHST can be summarized into three aspects. First, the procedure does not tell researchers what they want to know. In other words, NHST and scientific inference address different questions; 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 two incorrect widespread beliefs that the *p*-value is the probability that the null hypothesis is correct and the complement of a *p*-value is the probability that a significant result will be found in a replication (Kirk, 1996).

The second problem is that by adopting a fixed level of significance, researchers turn a continuum of uncertainty into an artificial dichotomous reject-or-do-not-reject decision. The use of 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 calculated *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. This dichotomous decision of statistically significant versus not statistically significant tells a researcher nothing about the practical significance or importance of a particular finding (Chow, 1988; Kirk, 1996; Shaver, 1993).

The third problem is that nearly all null hypotheses are false on a prior ground. The null hypothesis is always false, and a decision to reject it 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 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 always false, and therefore statistical significance testing becomes a vain effort of demonstrating what is already known (Kirk, 1996; Thompson, 1993; Vacha-Haase & Thompson, 2004).

These criticisms on 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, 1993; Schmidt & Hunter, 1995) have recommended complete elimination of significance testing, most scholars suggest that significance testing should be supplemented with or placed in the context of additional information, such as confidence intervals, odds ratio, and effect size (American Educational Research Association, 2006; Anderson, Burnham, & Thompson, 2000; Kirk, 1996, 2001; Fan, 2001; Mclean & Ernest, 1998; Vacha-Hasse & Thompson, 2004; Vaske, Gliner & Morgan, 2002; Snyder & Lawson, 1993; Thompson, 1996, 1997, 2000; Wilkinson & the APA Task Force on Statistical Inference, 1999). Reporting effect size is probably the most frequent recommendation (Ives, 2003).

Effect Size

What is 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; Thompson, 1998, 2000; Vacha-Haasse & Thompson, 2004). The 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 represented by Cohen's *d*, Glass's *g*, and Hedges' *g* (Cohen, 1988; Glass, 1976; Hedges, 1981); the latter is base on the proportion of variance accounted for or correlation between the independent variable and the dependent variable and represented by *R*-squared (R^2) and eta-squared (η^2) (Maxwell& Delaney, 1990; Snyder & Lawson, 1993; see Kirk 1996 for more details about the measures in each category).

The importance of effect size. Effect size values 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, which will help reduce the risk of statistically non-significant results because of inadequate sample size (Olejnik, 1984; Plucker, 1997). Second, it enables the other researchers and readers of the articles to have a clear understanding of the actual magnitude of treatment effect. 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 to one another. That is, they provide a statistical tool for meta-analysis that quantitatively synthesizes the effects across different studies.

The interpretation of effect size. The common practice in interpreting effect sizes is to use the benchmarks for "small", "medium" and "large" effects offered by Jacob Cohen in 1988. However, this is an extremely unfortunate practice and Cohen's benchmarks are not generally useful (Thompson, 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 relatively

established area of research, it is inappropriate to apply Cohen's guidelines blindly (Glass, McGaw, & Smith, 1981; Thompson, 2008). As proposed by Thompson (2008) the correct interpretation of effect size should focuses on the explicitly and directly comparing between effect size in new results and prior effect sizes in the related literature.

Statistical Significance vs. Practical Significance

Statistical significance refers to whether a result is due to chance or variability in the sample whereas practical significance refers to whether the result is useful in the real world. Practical significance makes it possible to make meaningful interpretations of research results and apply them to the real world.

A *p*-value helps to make judgment about the statistical significance of the results (Kirk, 1996; Snyder & Lawson, 1993; Volker, 2006). However, reporting *p*-value only is not enough to help readers to understand the practical significances of the study. First of all, *p*-values are confoundedly influenced by many factors and, therefore, *p*-values themselves cannot be used to decide the magnitude of the treatment effect (Falk & Greenbaum, 1995; Thompson, 1993; Thompson, 1999a). Moreover, *p*-values do not directly address the critical issue of result replicability; a smaller *p*-value does not imply greater confidence in the conclusion that sample results are replicable (Cohen, 1994; Thompson, 1996). Thus, it is recommended that significance testing should be reported and followed by effect sizes (Plucker, 1997; Robinson & Levin 1997; Thompson & Snyder, 1997) which can be used to make judgment about the practical significance of results. As Fan (2001) argued that *p*-value and effect size are two sides of one coin: they complement each other but they do not substitute for each other and, therefore, researchers should consider both sides. Since the purpose of research should be to measure the

magnitude of an effect rather than simply its statistical significance (Cohen, 1990), reporting and interpreting effect size is crucial.

Changing Publishing Policies

To respond to the criticisms about NHST and raise awareness of importance of effect size, journals and academic associations have changed their publication policies. In 1994, *Educational and Psychological Measurement*, the first journal requiring 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 American Psychological Association (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 was formed by APA in order to examine prevailing statistical practices, including statistical significance testing. The new recommendations emphasized that effect sizes should always be reported (Wilkinson & APA Task Force on Statistical Inference, 1999). The fifth edition of the Publication Manual of APA (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 American Educational Research Association (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).

Previous Studies of Effect Size Reporting Practices

Since the encouragement and urge to report effect size are nothing new, the effectiveness of these encouragements and urges is worthy of empirical examination. Review studies investigating the effect size reporting practices in journals and books are conducted under such environment. The earliest review of effect size reporting dated back to 1994. Dar, Serlin, and Omer (1994) reviewed 163 studies published in *Journal of Consulting and Clinical Psychology* in 1967-1968, 1977-1978, and 1987-1988 and found that, in general, effect sizes were increasingly being discussed in psychotherapy research. Reported effect size measures included squared multiple correlation coefficients and differences in proportions or means. However, in most cases, no effect size estimates were reported at all. The most recent review is conducted by Alhija and Levy (2007). Ninety-nine articles in 10 professional journals published in 2003 and 2004 were reviewed and no major differences existed between journals requiring effect size reporting and journals not requiring effect size reporting. The frequency of effect size reporting and interpreting depended on the statistical procedures; effect size measures in correlation and regression analysis are more frequently reported while effect size measures in *t*-test and regression are more frequently interpreted. Discrepancy existed between results based on statistical significance and practical significance in more than half of the articles that reported effect size measures and only few addressed the discrepancy. There are almost 20 other previous studies of effect size reporting, mostly in education and psychology area, and their major findings are summarized in Appendix A.

It was consistently found that effect size reporting practices varies across different NHST methods. Specifically, multivariate analyses are more likely to contain effect size than univariate (Alhija & Levy, 2006; Hutchins & Henson, 2002; Ives, 2003; Paul & Plucker, 2004). However, inconsistent conclusions were also identified. For example, Dunleavy et al.'s (2006) study found that variance-accounted-for effect sizes were typically omitted, but Kirk (1996) and McMillan, Lawson, Lewis, and Synder (2002) concluded that R^2 were the most often used effect size measures. Paul and Plucker (1997) did not see statistically different effect size reporting practice across six years, whereas Dar, Serlin, and Omer (1994) observed a general improvement across three ten-year periods.

Purposes of the Study

Though previous studies showed a promisingly increasing trend of effect size reporting practice, they have some limitations. First, the number of journals reviewed in most of the studies was four or less, a number that is probably not large enough to reveal the panorama and trends in education and psychology areas. Second, different studies have different focus. For example, Keselman et al. (1998) reviewed 17 journals which is a relatively large sample; but their focus was limited to ANOVA, MANOVA, and ANCOVA. Therefore, it is unable to show the whole picture of all different methods. Third, most of the 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 effect size interpreting practices need to be included in the review process. Fourth, effect size reporting and interpreting practices from 2005 to 2007 have to be reviewed and the inconsistent conclusions from previous studies mentioned eariler need to be further investigated.

Therefore, the purpose of the present study, following Alhija and Levy's (2007) review, is to review the effect size reporting and interpretation practices of quantitative studies published from 2005 to 2007 in a group of representative journals in education and psychology areas. Specifically, the following research questions are expected to be addressed in the present study:

1. What is the frequency of reporting effect size and what types of effect size measures are more frequently reported than others?
2. What is the frequency of interpreting effect size and what types of effect size measures are more frequently interpreted than others?
3. Is there any discrepancy between statistical significance and practical significance of the results? If yes, do the authors address the discrepancy?
4. Is there any difference between the effect size reporting and interpreting practices between different statistical methods, journal sponsors, and publication years?

Method

Data Source

Fourteen academic journals were selected for the present study and listed in Appendix B. The selected journals meet two major criteria: first, high proportion of quantitative empirical studies; second, frequently reviewed by previous review studies. To be more specific, among six journals published by AERA, included were 2 journals that mostly publish empirical studies. Six journals were from the APA online journal list by subject "Cognitive/Learning/Education". Six additional journals were taken from the original journal list of Alhija and Levy's (2007) study; and most of them were frequently reviewed in the previous studies, which makes it possible to compare the trends across different publication years. All the articles published from 2005 to 2007 were reviewed with the exclusion of book reviews, editorials, and journal announcements.

Quantitative studies and qualitative/quantitative mixed studies with NHST were included in this study.

Instrument

A 17-item checklist (see Appendix C) adopted from Alhija and Levy (2007) was used as the instrument for this present study. The following modifications were made on the original checklist: (a) "Year" was added as the second item because difference across publication year is one of the questions of the present study; (b) Item 3 of the original checklist, "type of journal (required/do not require reporting effect size in guidelines for authors)", was replaced by "Journal sponsor" since the main comparison was made between sponsors like association journals and independent journals; (c) Item 4 "Research topic" in the original checklist was replaced by the more specific item "Research questions"; (d) Items 6 and 7 "Participants" and "Research design and procedures" in the original checklist were deleted because they were irrelevant to effect size reporting and interpreting; (e) Item "Major analysis" was added; (f) Item 8 "Statistically significant/not significant" categorized all the test results into two opposite situations, significant or not, and neglected the circumstance that mixed results exist in testing with multiple groups or relations. Therefore, this item was improved as "Statistical result (significant/not significant/mixed)"; (g) Item 13 "The importance of reporting effect size values" in the original checklist was deleted; (h) Item 14 "The meaning of the effect size values in terms of research problems" in the original checklist was replaced by different wording "Is effect size interpreted?" and (i) Item "Is practical implication of the study discussed and how?" was added. Effect size reporting and interpreting are crucial for the studies with practical implication and therefore it is necessary to have this as a variable.

Procedures and Data Analysis Plan

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. If there is more than one major statistical method in one article, the first one was chosen. After all the eligible articles were reviewed, each variable was coded into different categories.

Since all the coded variables are nominal scales and do not produce numerical values that can be used to calculate means and variances, non-parametric test based on Chi-square statistic was applied to the analysis to address the research questions. Cramer's *V* was also reported 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 (Gravetter & Wallnau, 2006). As Cohen (1988) suggested, for Chi-square tests with degrees of freedom equal to 2, a value within the range of 0.07 to 0.21 is a small effect; a value within the range of 0.21 to 0.35 a medium effect and a value larger than 0.35 is large effect. However, as mentioned earlier in the study, the interpretation of effect size is context-dependent; it is problematic to apply Cohen's guidelines blindly. Therefore the present study reported Cramer's *V* without interpreting it.

Results and Discussion

Descriptive Statistics

Totally 1,581 empirical articles were published in 189 issues of the 14 journals from 2005 to 2007 and 78.6% 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 based on type of the journal, the main NHST method used and year published. Number of articles in each category was listed in Table

1. Of all the articles, 69.4% (*n* = 863) were published in APA journals while only 5.6% (*n* = 69) were published in AERA journals and 25% (*n* = 311) in independent journals. 75.5% of the articles (*n*=938) used general linear models as the main NHST methods. The numbers of articles across the three years do not vary very much; the percentages are 35.2%, 34.0% and 30.8% respectively.

Table 1 Descriptive Statistics: Number of Articles in Each Category

		n	%
Journal Type	AERA Journals	69	5.6
	APA Journals	863	69.4
	Independent Journals	311	25.0
Main NHST Method	Simple Tests	204	16.4
	General Linear Models	938	75.5
	Complex Models	101	8.1
Year Published	2005	438	35.2
	2006	422	34.0
	2007	383	30.8
	Total	1,243	

Reporting Effect Size

Table 2 summarizes the effect size reporting practices for each category. 49.1% (n = 610) of the 1,243 articles reported effect size. Within the journal type, the results $\chi^2_{(2)} = 84.695$, $p < .001$, Cramer's $V = .261$ indicate that there exist statistically significant differences between the three types of journals. Compared to the other two types of journals, AERA journals have the highest effect size reporting rate of 72.5%.

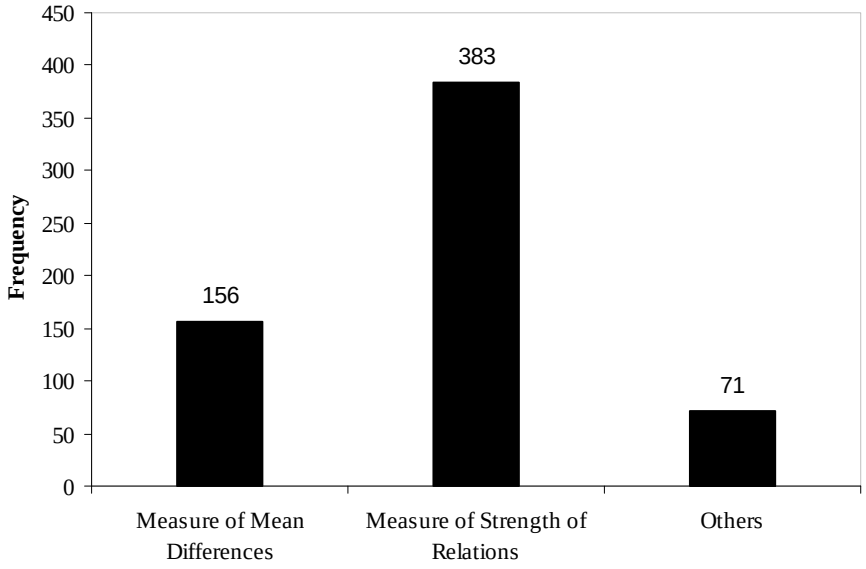
There is no statistical difference in effect size reporting within main NHST method type, $\chi^2_{(2)} = 4.947$, $p = .084$, Cramer's $V = .063$. The effect size reporting rate for complex models is the highest among the three types of main NHST while the rate for simple tests is the lowest. It is likely that researchers using complex models such as HLM and SEM have more advanced knowledge of statistics and therefore are more likely to report effect size; on the contrary, researchers using simple tests, especially the not so popular methods may not know which effect size measure to use or ignore the importance of reporting effect size. There is no statistical difference within publication year, $\chi^2_{(2)} = 5.659$, $p = .059$, Cramer's $V = .067$.

Of the 610 articles that reported effect size, the most frequently reported type of effect size measures is measure of strength of relations. See Figure 1 for details. This result is consistent with the findings in previous studies (e.g. Alhija & Levy, 2007; Hutchins & Henson, 2002; Kirk, 1996; McMillan, Lawson, Lewis, and Snyder, 2002); though Dunleavy et al. (2006) study 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 75.5% of the 1,243 articles (n = 938) used general linear models as the main NHST methods and 74.6% of the 610 articles that reported effect size (n = 455) used general linear models (cf. Table 1).

Table 2 Number of Articles that Reported Effect Size in Each Category

	Reported (%)	Not reported (%)	Total	$\chi 2$	*df*	*p*	Cramer's *V*
AERA Journals	50 (72.5)	19 (27.5)	69				
APA Journals	349 (40.4)	514 (59.6)	863	84.695	2	<.001	.261
Independent Journals	211 (67.8)	100 (32.2)	311				
2005	198 (45.2)	240 (54.8)	438				
2006	207 (49.1)	215 (50.9)	422	5.659	2	.059	.067
2007	205 (53.5)	178 (46.5)	383				
Simple Tests	95 (46.6)	109 (53.4)	204				
General Linear Models	455 (48.5)	483 (51.5)	938	4.947	2	.084	.063
Complex Models	60 (59.4)	41 (40.6)	101				
Total	610 (49.1)	633 (50.9)	1243				

Figure 1 Frequency of Effect Size Reporting for Different Measures



Interpreting Effect size

As discussed earlier in this study, applying Cohen's rules of thumb and 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). More advanced interpretation of effect size includes providing 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 (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. Figure 2 summarizes the number of articles that provided definition and justification of effect size measures. It is reasonable for the authors to assume that the readers have basic statistical knowledge and are aware of the definitions of the effect size measures, except for some uncommonly used measures. Therefore, in the present study, an article is identified as effect size interpreted as long as whether the effect is small, medium or large is indicated. This interpretation method is problematic, but it at least indicates the authors' awareness of the necessity of effect size interpreting.

Figure 2 Whether Definition and Justification of Effect Size Choice Were Provided

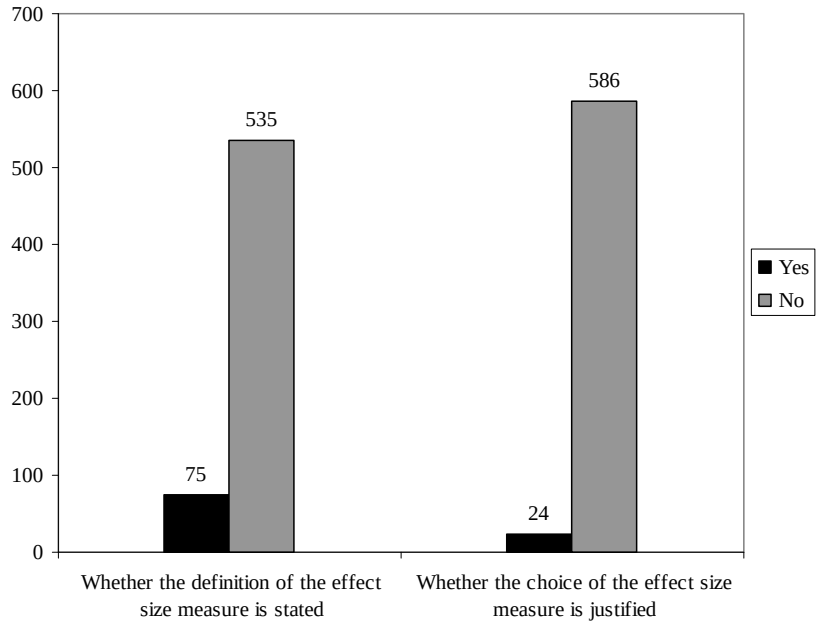


Table 3 summarizes the number of articles that interpreted effect size in each category and the Chi-square test results. Of the 610 articles that reported effect size, 56.7% (n = 346) contained interpretation of the values. There exist statistically significant differences within journal type, $\chi^2_{(2)} = 9.903$, $p = .007$, Cramer's $V = .127$. Independent journals have the highest rate of interpreting effect size, which is 64.5%, followed by AERA journals 62.0% and APA journals 51.3%. The overall rate is 56.7%, that is, 346 of the 610 articles that reported effect size also provide interpretation. There exist statistically significant differences within NHST method type, $\chi^2_{(2)} = 7.517$, $p = .023$, Cramer's $V = .111$. Similar to the finding in effect size reporting, articles that employed complex models are more likely to interpret effect size than others. Within the three years, there is no significant progress with regard to effect size interpreting as indicated by the Chi-square test result that $\chi^2_{(2)} = .427$, $p = .808$, Cramer's $V = .026$. In 2005, 58.6% of the articles contain interpretation of effect size; however, the rates for 2006 and 2007 are 55.6% and 56.1% respectively.

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 62.7% of the 346 interpreted effect size measures (n = 217). This is consistent to 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 to be interpreted in *t*-test and regression.

Table 3 Number of Articles that Interpreted Effect Size in Each Category

	Interpreted (%)	Not interpreted (%)	Total	χ^2	*df*	*p*	Cramer's *V*
AERA Journals	31 (62.0)	19 (38.0)	50				
APA Journals	179 (51.3)	170 (48.7)	349	9.903	2	.007	.127
Independent Journals	136 (64.5)	75 (35.5)	211				
Simple Tests	53 (55.8)	42 (44.2)	95				
General Linear Models	249 (54.7)	206 (45.3)	455	7.517	2	.023	.111
Complex Models	44 (73.3)	16 (26.7)	60				
2005	116 (58.6)	82 (41.4)	198				
2006	115 (55.6)	92 (44.4)	207	.427	2	.808	.026
2007	115 (56.1)	90 (43.9)	205				
Total	346 (56.7)	264 (43.3)	610				

Discrepancy between p- value and Effect Size

"Discrepancy" was defined as either statistically significant findings with small effect size or statistically not significant findings with medium to large effect size. The article was classified as "no discrepancy" if at least one of the effect size values consists with *p*-value. This loose classification criterion resulted in 69 out of 610 articles (11.3%) that have discrepancy between *p*-value and effect size. For example, in experiment 1 of Vachon, Tremblay and Jones' (2007) study, analysis of variance (ANOVA) produced two statistically nonsignificant interaction effects with *p*-value equal to .158 and .122 respectively; however, Cohen's *d* were .80 and .81 respectively, which are considered to be large effect by Cohen (1988).

Three Chi-Square tests were conducted to investigate the differences of discrepant findings across types of journals, types of NHST methods and three years. None of the results is statistically significant. The results were summarized in Table 4.

Table 4 Discrepancy between p-value and Effect Size in Each Category

	Discrepancy (%)	No discrepancy (%)	Total	χ^2	*df*	*p*	Cramer's *V*
AERA Journals	8 (16.0)	42 (84.0)	50				
APA Journals	36 (10.3)	313 (89.7)	349	1.502	2	.472	.050
Independent Journals	25 (11.8)	186 (88.2)	211				
Simple Tests	16 (16.8)	79 (83.2)	95				
General Linear Models	45 (9.9)	410 (90.1)	455	4.057	2	.132	.082
Complex Models	8 (13.3)	52 (86.7)	60				
2005	26 (13.1)	172 (86.9)	198				
2006	25 (12.1)	182 (87.9)	207	2.084	2	.353	.058
2007	18 (8.8)	187 (91.2)	205				
Total	69 (11.3)	541 (88.7)	610				

Whether the Discrepancy Was Address by the Authors

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 (2005) study, ANCOVA test did not produce a statistically significant result as indicated by F (2, 40) = 2.423, p = .102 but the effect size is .464 which is a large effect. In the discussion section the author explained that the absence of a robust difference is probably due to the small sample size, because the effect size was still large.

Of the 69 articles that have discrepant results based on *p*-value and effect size, 30.4% of them (n = 21) were addressed by the authors. Three Chi-square tests were conducted to investigate the differences between three types of journals, NHST methods and three years and the results are summarized in Table 5. Within three types of journals, the result is marginally significant, $\chi^2_{(2)} = 6.005$, p = .050, Cramer's V = .295. The tests across NHST methods and years are statistically non significant.

Table 5 Whether the Authors Address the Discrepancy between p-value and Effect Size

	Yes (%)	No (%)	Total	χ^2	*df*	*p*	Cramer's *V*
AERA Journals	1 (12.5)	7 (87.5)	8				
APA Journals	8 (22.2)	28 (77.8)	36	6.005	2	.050	.295
Independent Journals	12 (48.0)	13 (52.0)	25				
Simple Tests	5 (31.3)	11 (68.8)	16				
General Linear Models	14 (31.1)	31 (68.9)	45	.131	2	.937	.043
Complex Models	2 (25.0)	6 (75.0)	8				
2005	9 (34.6)	17 (65.4)	26				
2006	5 (20.0)	20 (80.0)	25	2.108	2	.349	.175
2007	7 (38.9)	11 (61.1)	18				
Total	21 (30.4)	48 (69.6)	69				

Conclusion and Implications

Reporting and interpreting effect size enables the readers to have a clear understanding of the actual magnitude of treatment effect. Because effect size is intended to be metric-free measures of the size of mean differences or strength of relations, they may be used to compare the results of different studies to one another. Previous studies found rates of effect size reporting ranging from 1% (Meline & Schmitt, 1997) to 87% (Thompson, 1999a) and it is 49.1% in the present study. The rate of effect size interpreting is about 40% in Alhija and Levy (2007) study, 50% in Meline and Wang (2004) study, 88% in Hutchins and Henson (2002) study, and 56.7% in the present study. Because the reviewed journals and review criteria are different between the studies, especially that some previous studies used very small sample sizes, e.g. Hutchins and Henson (2002) study used a sample size of 14 articles and Thompson and Snyder (1998) study used a sample size of 22 articles, it is very difficult to compare those results to the present study. However, within the context of the present study, an overall rate of 49.1% for effect size reporting and 56.7% for effect size interpreting is still far from satisfactory.

The present study shows that effect size reporting practice differs between journal types but does not differ between different types of NHST methods, which contradict with previous studies (Alhija and Levy, 2007; Dunleavy et al, 2006; Hutchins & Henson, 2002; Ives, 2003; Paul & Plucker, 2004). As far as effect size interpreting practice is concerned, it differs between 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 is low and suggests that researchers should

pay attention to this question when analyzing the data and writing the report. None of the four tests about the time effect are significant and implies that there is no statistically significant improvement on effect size reporting and interpreting practice from 2005 to 2007.

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) mentioned that this may be due to the fact that those measures are usually produced automatically through the significance testing procedure, e.g. R^2 in regression family, and cannot be totally interpreted as high awareness of the importance of reporting and interpreting effect size measures. The popularity of measures of strength of relations is consistent with the fact that general linear models are the most frequently used NHST methods in the present study. Therefore this result should be interpreted with caution.

The result that 11.3% of the 610 articles reported effect size has discrepant results based on *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 have discrepant results, only 30.4% of them (n = 21) contain the possible reasons for the discrepancy. This low percentage is consistent with the findings in Alhija and Levy (2007) study. Many researchers tend to ignore the meaning and importance of effect size measures in the context of their research or check the quality of their results. In a majority of the 21 articles, the discrepancy addressed by the authors in a very sloppy way 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 as the Simard and Nielsen (2005) study quoted earlier did. 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 purpose of the present study is to motivate researchers to pay close attention to reporting and interpreting effect size measures. Based on the results and the possible reasons discussed earlier, it is necessary for the academic journals, leading scholars, and academic associations to continue to urge the improvement of effect size reporting and interpreting practices. Considering the publishing lag of academic journals, the results in the present study may not be able to reflect the impact of APA requirements and AERA 2006 Standards. Because of the relativity of effect size values (Cohen, 1988; Kirk, 1996; Thompson, 2000), authors are strongly recommended to report and interpret effect size measures not only in the measure themselves but also in the specific context of their research questions.

Researchers' resistance to report effect size may be partially explained by some combination of confusion and desperation about NHST and effect size (Thompson, 1999b). One of the sources of researchers' confusion may come from textbooks. Two review studies on statistics books and textbooks (Capraro & Capraro, 2002; Curtis & Araki, 2002) suggested that insufficient attention was given to effect size compared with NHST; effect size parameters and statistics were not distinguished; how to calculate and interpret effect size statistics were not agreed. Statistics textbooks are the tools of researchers, students, and future researchers; those problems with effect size in textbooks may affect their practice in research. Articles about how to understand different types of effect size measures contributed to the literature to alleviate researchers' confusion about effect size (e.g., Cromwell , 2001; Glass, 1976; Hojat and Xu, 2004; Kirk, 1996; Lakshmi, 2000; Mahadevan, 2000; Robey, 2004; Snyder & Lawson, 1993; Smithson, 2001; Trusty, Thompson & Petrocelli, 2004; Vacha-Haase, & Thompson, 2004; Volker, 2006)

As an aid to improving the effect size reporting and interpreting practice, software companies are recommended to make the calculation of effect size measures as a default in NHST. When writing manuscripts, researchers should not assume that the readers know everything; the important definitions, justifications or interpretation of the effect size measures should not be omitted. Thompson (2008) provided a good guidelines on how to interpreting effect sizes from the methodological perspective; content experts are suggested to provide criteria for interpreting effect size values in specific research areas so that researchers will not blindly follow Cohen's benchmarks. The editors of the journals can play the role of a gate keeper. Effect size reporting and interpreting practices can be improved significantly with the joint efforts from different parties.

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(6),* 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.

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.

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

Capraro, R. M., & Capraro, M. M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement, 62(5),* 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.

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.

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.

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(4),* 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. *The Journal of Educational Research, 94(4),* 275-283.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5(10),* 3-8.

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: Thomson Wadsworth.

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

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

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.

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

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(6),* 490-504.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication, 23(3),* 251-266.

Keselman, H. J., Huberty, C. J., Lix, L.M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68(3),* 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(2),* 213-218.

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, January 28-29.

Lance, T. S., & Vacha-Hasse, 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). The effect size statistic: overview of various choices. Paper presented at the annual meeting of the southwest educational research association, Dallas, TX, January 28-29.

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(3),* 231–257.

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.

McLean, E. J. & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools. 5(2),* 15-22.

Meline, T., & Schmitt, J. F. (1997). Case studies for evaluating statistical significance in group designs. *American Journal of Speech-Language Pathology, 6(1),* 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.

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

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(2),* 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*. A 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(5),* 21–27.

Robinson, D. H., & Wainer, H. (2002). On the past and future of null hypothesis significance testing. *Journal of Wildlife Management, 66(2),* 263–271.

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 & the Health Professions, *18(4),* 408–427.

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

Simard, V. & Nielsen, T. A. (2005). Sleep paralysis–associated sensed presence as a possible manifestation of social anxiety. *Dreaming, 15(4),* 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(4),* 335-348.

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(2),* 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26(5),* 29-32.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of possible future. *Research in the Schools, 5(2),* 33-38.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9(2),* 167-183.

Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65(3),* 329-337.

Thompson, B. (2000). "Statistical", "practical", and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80,* 64-71.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In Osborne, J. W. (Eds.) *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, 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., 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(3),* 413-425.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51(4),* 473-481.

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(1),* 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(1),* 46-57.

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(2),* 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*(6), 653-672.

Ward, R. M. (2002). *Highly significant findings in psychology: A power and effect size survey*. 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.

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

Appendix A: List of Previous Review Studies

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Alhija & Levy, 2007	Contemporary Educational Psychology Journal of Learning Disabilities Exceptional Children Early Childhood Research Quarterly Journal of Experimental Education Journal of Learning Disabilities Journal of Educational Psychology Learning Disabilities Research & Practices Journal of Special Education Infant and Child Development Educational Research	2003-2004	99	The frequency of effect size reporting and interpreting depended on the statistical procedures; effect size values were more frequently to be reported in correlation and regression while more frequently to be interpreted in *t*-test and regression. Discrepancy existed between results based on statistical significance and practical significance in more than half articles that reporting effect size measures and only few addressed the discrepancy.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Dar, Serlin & Omer, 1994	Journal of Consulting and Clinical Psychology	1967 - 1968 1977 - 1978 1987 - 1988	163	In general, effect size is increasingly being discussed in psychotherapy research. Whereas only 14.8% of studies in 60s explicitly reported effect size, this proportion grew to 29.7% in the 70s and by 80s 61.4% of the studies included effect size measures (correlation between decade and proportion of studies reporting effect size was .33, $p < .005$). ES estimates included squared multiple correlation coefficients and differences in proportions or means. However, in most cases, no effect size estimates were reported at all. Specially, no measures of effect size (i.e. eta or omega squared or other measures of the percent of variance accounted for by the independent variables) were ever reported in the context of an ANOVA.
Dunleavy et al., 2006	Journal of Applied PsychologyJournal of Educational PsychologyJournal of Personality and Social PsychologyJournal of Educational Psychology, Learning and Memory	2002-2003	736	Overall 62.5% of all articles reported effect sizes but the effect size reporting varied by journals. Univariate analyses testing mean differences had the greatest number of omitted effect size. Variance-accounted-for statistics were typically omitted. JAP and JPSYCH reported more effect sizes than the others.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Hutchins & Henson, 2002	Communication Education	2000	14	Eight articles (57%) reported effect sizes for their statistically significant results. The effect sizes used most were eta squared (50%) and *r*-squared (30%) with omega squared and Cohen's *d* used in the remaining studies. Of 8 articles that reported effects, 7 (88%) interpreted the effect size measures in their discussion. All emphasized how important and thus practically significant results were, beyond *p*-value.
Ives, 2003	Journal of Learning Disabilities Learning Disabilities Research & Practice Learning Disability Quarterly	1990-1999	526	A total of 526 quantitative studies with an overall effect size reporting rate of about 25%. The overall effect size reporting rates across all 10 years were 21% for univariate studies and 43% for multivariate studies.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Keselman et al., 1998	American Educational Research JournalChild DevelopmentCognition and InstructionContemporary Educational PsychologyDevelopmental PsychologyEducational Technology, Research and DevelopmentJournal of Applied PsychologyJournal for Research in Mathematics EducationJournal of Counseling PsychologyJournal of Educational Computing ResearchJournal of Educational PsychologyJournal of Experimental Child PsychologyJournal of Experimental EducationJournal of Personality and Social PsychologyJournal of Reading BehaviorReading Research QuarterlySociology of Education	1994 or 1995	411	For Between-Subjects Univariate Designs: The issue of power and/or effect size calculations arose in only 10 articles (16.1%). Effect sizes were calculated in six of these articles, but the statistic used was not routinely reported, and main effects were more often of interest than interactions.For Between-Subjects Multivariate Designs: Effect size index values were reported in only 8 of the 79 articles. Seven studies used univariate indexes, and one study reported multivariate eta-squared values. For Repeated Measures Designs: Issues of statistical power/effect size were considered in 20 of the 226 articles (8.8%) in the database. In 16 of these articles, effect sizes were calculated, with the most common measure being Cohen's (1988) d statistic. In three articles, the authors mentioned that statistically significant findings may not have been revealed because of potentially low power, but no assessments of power were actually performed. For Covariance Designs: only 15 studies reported adjusted means (it was assumed that reported means were unadjusted unless explicitly stated); 11 studies provided some index of effect size, with standardized mean difference being the most popular (seven studies); and none of the studies examined reported results in terms of confidence intervals

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Kirk, 1996	Journal of Applied Psychology Journal of Educational Psychology Journal of Experimental Psychology, Learning & Memory Journal of Personality and Social Psychology	1995	391	In numbers of measures of effect magnitude, there is a considerable variability among the journals: 77% of the articles in JAP contained one or more measures of effect m while JEP was only 12% (due to the testing procedures they use. The former are more likely to use regression and correlation while the latter use analysis of variance). The 3 most frequently used inferential procedures were analysis of variance, the t test for means, regression analysis. R squared is the most popular measure of association strength whereas bias-corrected counterparts of R squared have been minimally reported.
Lance & Vacha-Hasse, 1998	The Counseling Psychologist	1995-1996		40.5% of the reviewed articles reported effect size (cf. Vacha-Haase et al. 2000)

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
McMillan, Lawson, Lewis & Synder, 2002	Journal of Educational PsychologyJournal of Educational ResearchContemporary Educational PsychologyJournal of Experimental Education	1997-2000	508	Of 508 articles classified as quantitative or mixed methods, 148 mentioned or calculated effect size, but only 82 articles included a calculation of effect size and at least limited discussion of magnitude or practical significance. Only 30 articles contained a calculated effect size and extensive discussion about effect size or magnitude. *R*-squared was the most used association statistic, followed by *r*-squared and eta squared, while Cohen's d was the most common difference statistic used.
Meline & Schmitt, 1997	American Journal of Speech-Language Pathology American Journal of Audiology Language, Speech, and Hearing Services in Schools Journal of Speech, Language, and Hearing Research	1990-1994	411	411 research articles in ASHA journals were examined and it was found that effect size was reported in only 5 of 411 articles.
Meline & Wang, 2004	American Journal of Speech-Language Pathology American Journal of Audiology Language, Speech, and Hearing Services in Schools Journal of Speech, Language, and Hearing Research	1999-2003	433	Effect-size statistics were reported in 27.7% of the articles overall, but results for the individual journals varied widely and ranged from 72% (LSHSS) to 13% (AJA). Although many authors reported effect size, nearly half of the authors did not interpret their effect-size results.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Paul & Plucker, 2004	Journal for the Education of the Gifted Roeper Review Gifted Child Quarterly	1995-2000	723	Overall about 28.9% of the quantitative research blocks contained effect size estimates. No statistically significant differences between the journals on the reporting of effect size estimates. No statistically significant differences in effect size reporting across the three time periods (1995- 1996, 1997-1998, and 1999-2000). Multivariate analysis (52.2%) contained effect size estimates more often than did univariate research blocks (17.9%).
Plucker, 1997	Gifted Child QuarterlyJournal for the Education of the GiftedRoeper Review40 studies selected from ERIC CD-ROM	1992-1995		The lack of effect size information is consistent across journals, and there was no significant difference among journals. The percentage of effect size blocks in the gifted journals was similar to that in the non-gifted journal. There was neither statistically nor practically significant differences between the time periods (1992-1993 and 1994-1995) with respect to effect size reporting. Results indicate that multivariate blocks included effect sizes information more frequently than univariate blocks.
Snyder & Thompson, 1998	School Psychology Quarterly	1990-1996	35	19 articles reported various magnitudes of effect indices but few interpreted them. Only 2 of the 35 articles invoked an “internal” replicability analysis.Almost all authors who failed to reject their null hypotheses did not conduct power analyses.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Thompson, 1999	Exceptional Children	1996-1998	23	Effect sizes were not reported in 20 of 23 quantitative articles. Effect size was calculated and a medium effect was identified for both the mean difference measures and variance-accounted-for statistics.
Thompson & Snyder, 1997	Journal of Experimental Education	1994 - 1996	22	In 4 articles, effect sizes were the focus of result presentation and interpretation. In articles in which effect-size estimates were reported, eta2, Cohen's d, and the conversion of effects to r were the vehicles for reporting these estimates.
Vacha-Haase & Ness, 1999	Professional Psychology: Research and Practice	1990-1997	204	Of the 265 quantitative research articles, 204(77.0%) used statistical significance testing. Authors often describe their results as "significant" rather than using the complete phrase to describe this concept. Less than 20% of the authors correctly used the term "statistical significance". The majority of authors made no mention of the effect size.

Authors and Year	Journals Covered	Years Covered	Sample Size	Major Findings
Vacha-Haase & Nilsson, 1998	Measurement and Evaluation in Counseling and Development	1990-1996	83	A minority of the published articles presented their statistically significant results by indexing results to sample size (7.3%) or reported effect sizes (35.3%). Regarding placing test results in a sample-size or an effect-size context, as encouraged in MECD's guidelines for authors, only a minority of the reviewed articles presented their statistically significant results in context of the sample size index (7.3%) and calculated effect size (35.3%).
Vacha-Haase et al., 2000	Journal of Counseling Psychology Psychology and Aging	1995-1997	277	All the articles in 1967, 1977, 1987, 1990-1997 were reviewed and the results indicated that authors usually merely describe their results as “significant”, rather than as “statistically significant”. As regards effect size reporting, about half of the Psychology and Aging articles in which statistical tests were used reported at least one effect size the frequency in Journal of Counseling Psychology. There has been some increased effect-size reporting in recent years, notably in the last three years we studied, especially in comparison with 1967, 1977 and 1987.
Ward, 2002	Journal of Consulting and Clinical PsychologyJournal of Personality and Social PsychologyJournal of Abnormal Psychology	2000	287	About 7% of studies estimated or discussed statistical power, and about 30% calculate effect size measures. These numbers were far below the desired level of mandatory reporting of these measures.

Appendix B: A List of Reviewed Journals and Their Sponsors

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

Appendix C: Revised 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?	☐
9	Are effect size reported?	☐
10	Are effect size reported also for not significant results?	☐
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?	☐
14	Is the effect size interpreted?	☐
15	Is there a discrepancy between conclusions based on statistical as opposed to practical significance?	☐
16	If such a discrepancy exists, has the author address it?	☐
17	Is practical implication of the study discussed?	☐