Article

HEALTH EDUCATION & BEHAVIOR

Health Education & Behavior
2016, Vol. 43(5) 518–527

DOI: 10.1177/1090198116669521
heb.sagepub.com
SAGE

Failure to Report Effect Sizes: The Handling of Quantitative Results in Published Health Education and Behavior Research

Adam E. Barry, PhD[1], Leigh E. Szucs, MEd, CHES[1], Jovanni V. Reyes, MS, CHES[1], Qian Ji, MS[1], Kelly L. Wilson, PhD, MCHES[1], and Bruce Thompson, EdD[1]

Abstract

Given the American Psychological Association's *strong* recommendation to always report effect sizes in research, scholars have a responsibility to provide complete information regarding their findings. The purposes of this study were to (a) determine the frequencies with which different effect sizes were reported in published, peer-reviewed articles in health education, promotion, and behavior journals and (b) discuss implications for reporting effect size in social science research. Across a 4-year time period (2010-2013), 1,950 peer-reviewed published articles were examined from the following six health education and behavior journals: *American Journal of Health Behavior*, *American Journal of Health Promotion*, *Health Education & Behavior*, *Health Education Research*, *Journal of American College Health*, and *Journal of School Health*. Quantitative features from eligible manuscripts were documented using Qualtrics online survey software. Of the 1,245 articles in the final sample that reported quantitative data analyses, approximately 47.9% (n = 597) of the articles reported an effect size. While 16 unique types of effect size were reported across all included journals, many of the effect sizes were reported with little frequency across most journals. Overall, odds ratio/adjusted odds ratio (n = 340, 50.1%), Pearson r/r^2 (n = 162, 23.8%), and eta squared/partial eta squared (n = 46, 7.2%) accounted for the most frequently used effect size. Quality research practice requires both testing statistical significance and reporting effect size. However, our study shows that a substantial portion of published literature in health education and behavior lacks consistent reporting of effect size.

Keywords

APA reporting guidelines, Cohen's effect size benchmarks, effect size, research, statistical best practice, statistical significance testing

The use, interpretation, and reporting of statistical significance tests have been the focus of contentious debates across a wide array of disciplines (Anderson, Burnham, & Thompson, 2000; Cohen, 1994; Fidler, Cumming, Burgman, & Thomason, 2004; Harlow, Mulaik, & Steiger, 1997; Krueger, 2001; Meehl, 1978; Nickerson, 2000; Thompson, 1996; Trafimow & Rice, 2009; Vacha-Haase, 2001). In fact, some scholars have asserted that severe deficiencies associated with null hypothesis statistical significance testing (NHSST) "retards the development of cumulative knowledge" (Schmidt, 1996, p. 115). Due to the limitations associated with statistical significance testing, some have gone so far as to propose abandoning statistical significance testing entirely (see Carver, 1993; Schmidt, 1992, 1996). Recently, the journal *Basic and Applied Social Psychology* banned the reporting of NHSST, stating,

> We believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research. We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking. (Trafimow & Marks, 2015, pp. 1-2)

Recognizing that a universal banning of statistical significance testing is probably unlikely, other scholars instead have proposed that testing of statistical significance can be

[1]Texas A&M University, College Station, TX, USA

Corresponding Author:
Adam E. Barry, Department of Health and Kinesiology, Texas A&M University, TAMU 4243, College Station, TX 77843, USA.
Email: aebarry@tamu.edu

employed and properly interpreted *if* researchers contextualize their results by also reporting and interpreting associated effect sizes (Buhi, 2005; Cohen, 1994; Colliver, 2002; Fan, 2001; Kirk, 1996; Paul & Plucker, 2004; Thompson, 1999c; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000). Maher, Markey, and Ebert-May (2013) argued that effect sizes represent "the other half of the story" (p. 345). Despite promoting for decades the reporting of effect sizes to supplement NHSST (Kirk, 1996; O'Fallon et al., 1978), fields ranging from psychology to special education exhibit poor effect size reporting practices (cf. Finch, Cumming, & Thomason, 2001; Kirk, 2001; Thompson, 1999d). Additionally, the guiding principles for reporting statistical analyses in biomedical science explicitly state that investigators should "describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results" (Lang & Altman, 2013, p. 3).

To date, there have been limited examinations of effect size reporting practices in the published health education and behavior literature. We believe such an examination is critical given that the use of statistical testing without associated "statistical literacy" could lead a field of study to become less scientific (Westover, Westover, & Bianchi, 2011). In an attempt to contribute to the statistical literacy of health educators, the present investigation (a) outlines some of the issues surrounding the use of NHSST; (b) examines the effect size reporting practices of researchers publishing in the field of health education, promotion, and behavior; and (c) emphasizes the importance of not using rigid effect sizes benchmarks for interpretation purposes.

Cohen (1988) proposed benchmarks for what might be deemed "small," "medium," and "large" effect sizes (e.g., $d = |0.2|$ = small, $d = |0.5|$ = medium, $d = |0.8|$ = large; $\eta^2 = 2\%$ = small, $\eta^2 = 10\%$ = medium, $\eta^2 = 25\%$ = large) as a means to contextualize/evaluate effect size indices. However, scholars such as Thompson (2004) and Prentice and Miller (1992) have argued against the rigid use of Cohen's fixed standards, asserting the arbitrary "one-size-fits-all" rule hinders the correct interpretation of effect size. Cohen, himself, stated that the benchmarks were subjective and created "with much diffidence, qualifications, and invitations *not to employ them* [italics added] if possible" (p. 12). Others have asserted that such indices should be considered only a "rough guide" (Howell, 2002, p. 206; Richardson, 2011) for interpretation.

Null Hypothesis Statistical Significance Testing

First, we briefly outline some limitations of NHSST. Trafimow and Rice (2009) succinctly characterized NHSST by saying that NHSST

> requires that the researcher propose a null hypothesis and an alternative hypothesis, collect data, and use the data to compute the probability of obtaining a finding as extreme or more extreme than the one actually obtained, given that the null hypothesis is true. If this probability is low (e.g., $p < .05$), then the researcher rejects the null hypothesis in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected. (p. 261)

Moreover, it is important that researchers understand what a "*p*" value really constitutes. A *p* value is merely the calculated probability of the sample findings *given (a) the null hypothesis is assumed true in the population and (b) taking into consideration the sample size* (McLean & Ernest, 1998; Thompson, 2006a). Interested readers are directed to Goodman (2008) for a list of 12 common misconceptions and misuses of *p* values.

Limitations of NHSST

Several limitations of statistical significance testing can be cited. First, simply because a sample finding is unlikely or improbable does not necessarily mean that the finding is important (Thompson, 1996, 2006a). Moreover, null hypotheses are almost always false on a priori grounds (Johnson, 1995), given that parameter values for different populations for continuous variables almost certainly differ to at least some large number of decimal places (Meehl, 1978). For example, all men and all women in the population of adults may be very identical in their mean IQ scores, but there is likely some difference in the two IQ means at some values (probably) far to the right of the decimal point. Thus, the decision to reject the 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" (Kirk, 1996, p. 747).

Second, NHSST is explicitly tied to sample size, such that even minor/trivial differences can become statistically significant given a large enough sample (Fan, 2001; Thompson, 1992, 1993, 1999b). Third, statistical significance does not provide insight regarding the replicability of sample results (Lang & Altman, 2013; Thompson, 1996, 2006a). For more detailed treatments on the limitations associated with NHSST, we refer interested readers to Anderson et al. (2000), Carver (1993), Cohen (1994), Frick (1996), and Thompson (1996, 2004).

Why Effect Sizes Are Important and Informative

Unlike *p* values generated from significance tests, effect sizes directly reflect the strength of the relationship between variables or the magnitude of intervention effects (Plucker, 1997; Thompson, 1999a), something statistical significance is unable to do (Colliver, 2002). The appealing feature of effect sizes is that they can be used to "report and interpret results in ways that are trustworthy, usable and, especially, are accessible to both scholars and practitioners" (Thompson, 1999b, p. 332). Grissom and Kim (2005) noted that readers "have a right to see estimates of effect sizes. Some might even argue

that not reporting such estimates in an understandable manner . . . may be like withholding evidence" (p. 5).

Moreover, reporting effect size is beneficial because it can facilitate both "meta-analytic thinking" and meta-analyses themselves within and across a published literature (Lang & Altman, 2013; Thompson, 1999a). In fact, Vacha-Haase (2001) contended that outside of a true replication study, the ability to compare effect sizes across published research is a "next-best alternative" (p. 220).

Method

Across a 4-year time period (2010-2013), we examined a total of 1,950 peer-reviewed published articles from six prominent health education and behavior journals: *American Journal of Health Behavior* (*AJHB*), *American Journal of Health Promotion*, *Health Education & Behavior* (*HE&B*), *Health Education Research, Journal of American College Health*, and *Journal of School Health*. A 4-year publication span was selected given that previous investigations have used this length of time to present and establish current reporting practices in the field of health education and behavior (see Barry, 2005; Barry, Chaney, Piazza-Gardner, & Chavarria, 2014; Smith, 2009). These specific journals were selected for several reasons. First, several of the journals represent the "flagship" publication of the major professional societies, including the American College Health Association (*Journal of American College Health*), American School Health Association (*Journal of School Health*), and Society for Public Health Education (*HE&B*)—all of which are members of the Coalition of National Health Education Organizations. Second, all six journals are peer-reviewed and indexed in the same Social Sciences Citation Index Category of Public, Environmental and Occupational Health, with 2015 impact factors (Thomson Reuters, 2016) ranging from a high of 2.312 (*HE&B*) to a low of 1.27 (*AJHB*). Finally, these journals have been included in previous investigations that have sought to assess the reporting practices of researchers who have published in the health education and behavior literature (Barry, 2005; Barry et al., 2014).

Inclusion and Exclusion Criteria

To be included in the investigation, articles had to meet the following criteria: (a) report quantitative data, (b) be published between the years of 2010-2013, and (c) appear in one of the aforementioned health education, promotion, and behavior journals. Qualitative research, duplicate articles, letters to the editor, commentaries, policy/practice recommendations, and conceptual and/or theoretical pieces were excluded. Additionally, articles that we were unable to retrieve because of incorrect cataloging in the database (n = 32) were excluded from the investigation. Mixed-method studies (i.e., analyses that used both quantitative and qualitative methods) were included; however, only the quantitative portions of these reports were analyzed.

Procedure

To capture an inclusive representation of all published literature, no formal search algorithms were employed; rather, the authors examined every article from each journal issue/volume published between 2010 and 2013. Figure 1 summarizes the article selection process that was followed (Moher, Liberati, Tetzlaff, Altman, & the PRISMA Group, 2009). Published articles meeting the inclusion criteria were subsequently screened to determine whether they reported either a standardized differences effect size, a variance-accounted-for effect size, or a "miscellaneous" effect size, and also the manner in which the reported effect sizes were interpreted was examined. A total of 1,245 published articles constituted the final sample for the investigation. This sample is a comprehensive illustration of the health education and behavior literature, representing a total of 24 volumes (4 × 6) across a 4-year time span and six journals.

Analyses

Data from each eligible article were documented using Qualtrics online survey software. In addition to documenting pertinent bibliographical information (authors, year of publication, journal source, volume, pages), we examined (a) whether the author(s) reported a measure of effect size and (b) which effect size was reported, if one was reported. Thompson (1999a), Kirk (1996), and Onwuegbuzie, Levin, and Leech (2003) have categorized effect size measures into three broad classes: variance-accounted-for measures (e.g., R^2, η^2), standardized differences (e.g., Cohen's d, Glass's delta), and "other/miscellaneous" effect indices (e.g., odds ratio, relative risk). For this investigation, we included all types of effect sizes that were reported in the published articles.

Results

Reporting of Effect Sizes

Table 1 shows the overall number of articles that reported a type of effect size across the six journals. Of the 1,245 articles in the final sample, 47.9% (n = 597) reported one of the aforementioned types of effect size. Thus, less than half of all health education and behavior research published during the 4-year time period in the six journals reviewed adheres to best-practice statistical reporting. The annual reporting percentages ranged from a low of 30.7% (*American Journal of Health Promotion*, in 2013) to a high of 67.1% (*AJHB*, in 2011). Table 2 shows the overall frequencies of the types of effect size reported in three categories (i.e., variance-accounted-for, standardized-differences, and "other" effect indices) across all the published articles (n = 597) that reported effect sizes. Table 2 shows that approximately 16 types of effect sizes were reported across the six journals, resulting in a total of n = 675 individual effect sizes recorded.

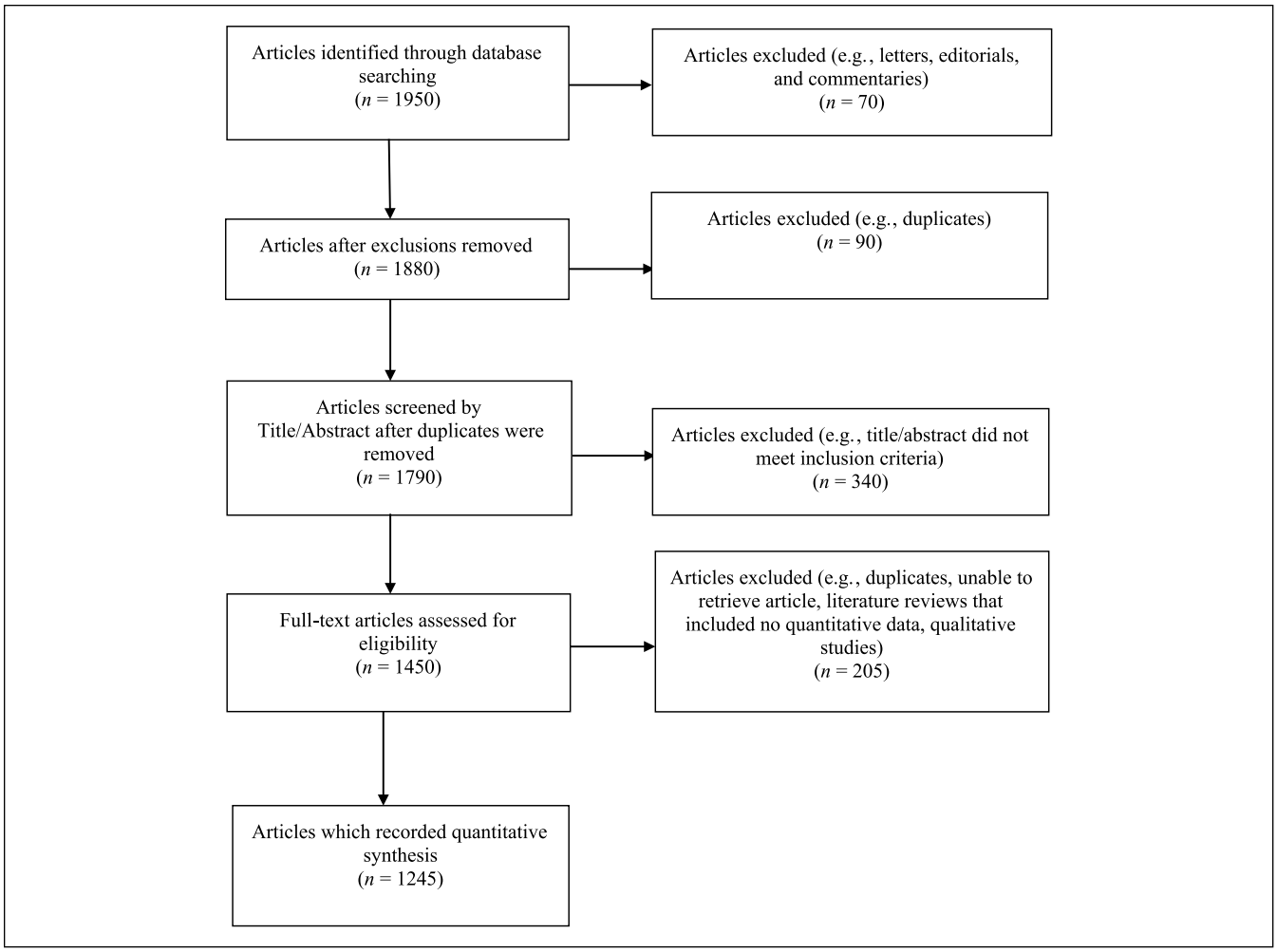

Figure 1. Flow of article selection process to identify the final sample of articles reporting quantitative analyses that were included in this investigation.
Note. Adapted from Moher, Liberati, Tetzlaff, Altman, and the PRISMA Group (2009).

It is noteworthy that the most frequently reported types of effect size included "other measures" of effect size, such as the odds ratio/adjusted odds ratio, followed by "variance-accounted-for" effect sizes, such as the Pearson r/r^2, and eta squared/partial eta squared (η^2). The journal-specific percentages reported in Table 2 were calculated by dividing the number of individual effect size types reported (e.g., *AJHB*, $r/r^2 = 40$) by the number of journal articles that reported an effect size (e.g., *AJHB*, $n = 125$, as shown in Table 1). For example, 32% of the articles reporting an effect size in *AJHB* were the r/r^2 effect size type.

While there are no "clear-cut choices" of what would constitute an optimal effect size in a given inquiry (Thompson, 1999a)—variance-accounted-for measures, standardized differences, or other indices—it is noteworthy that standardized differences effect sizes convert all effects to the metric of standard deviations and thus average differences can be meaningfully compared across samples or variables having initially different standard deviations (Thompson, 1999b). Of the total number of reported effect sizes ($n = 673$), 39.1% ($n = 263$) were "variance-accounted-for," 8.6% ($n = 58$) were standardized differences, and 52.3% ($n = 352$) were "other" or miscellaneous indices. Figure 2 illustrates the variance-accounted-for effect size reporting trends and Figure 3 displays the standardized differences effect size reporting trends between 2010 and 2013 across each of the six journals investigated in this study.

Interpretation of Effect Sizes

To provide further context to the manner in which effect sizes are used in published health education and behavior research, we examined how a standardized difference effect size (Cohen's *d*) and a variance-accounted-for effect size (eta-squared/partial eta-squared) were interpreted by researchers. A total of 41 articles reported a Cohen's *d* standardized differences effect size. Of those, 17 articles cited the "small," "medium," and "large" benchmarks proposed by Cohen (1988, 1992), notwithstanding his own admonitions to not use his

Table 1. Annual Effect Size (ES) Reporting Frequency and Percentages for Six Journals That Publish Health Education and Behavior Research, 2010-2013.

Journal title and years	Annual ES reporting, %	Not reporting, *n*	Frequency of ES reporting, *n*	Total *n*
American Journal of Health Behavior		122	125	247
2010	42.8	36	27	63
2011	67.1	21	43	64
2012	46.2	36	31	67
2013	45.2	29	24	53
American Journal of Health Promotion		109	87	196
2010	45.2	23	19	42
2011	52.9	32	36	68
2012	42.5	27	20	47
2013	30.7	27	12	39
Health Education & Behavior		80	77	157
2010	47.3	20	18	38
2011	48.5	18	17	35
2012	42.5	23	17	40
2013	56.8	19	25	44
Health Education Research		105	109	214
2010	53.8	30	35	65
2011	53.4	27	31	58
2012	54.1	22	26	48
2013	39.5	26	17	43
Journal of American College Health		95	92	187
2010	33.3	20	10	30
2011	52.6	36	40	76
2012	46.5	31	27	58
2013	65.2	8	15	23
Journal of School Health		137	107	244
2010	37	34	20	54
2011	38.2	42	26	68
2012	51.9	25	27	52
2013	49.2	36	34	69
Total journal articles sample	47.9	648	597	1,245

benchmarks, if possible. Three articles reported their own benchmark levels to interpret their effect sizes.

A total of 46 articles reported an eta squared/partial eta squared (η^2) effect size. Similar to the Cohen's *d* interpretive results noted above, 19 articles used Cohen's (1988, 1992) *d* benchmarks of "small," "medium," and "large" to interpret their effects, while 3 articles referenced Cohen's (1988) partial eta-squared benchmarks to interpret size of effect. The vast majority of the articles reporting Cohen's *d* ($n = 21$) and eta-squared/partial eta-squared ($n = 27$) values neglected to provide any interpretation or discussion about the magnitudes of their effect; the effect sizes were simply reported without any interpretation.

Discussion

Given that *p* values represent confounded indices (subject to the joint influences of both sample sizes and effect sizes) and ultimately do not give insights into the relative importance of results, scholars have proposed the reporting of effect size as one strategy for the improvement of practice (Thompson, 1999a). The American Psychological Association (APA, 1994) has not only "encouraged" (p. 18) the reporting of effect sizes but also gone so far as to assert that researchers should "always provide some effect size estimate when reporting a *p* value" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). Moreover, recent recommendations from the APA's Publications and Communications Board and Working Group on Journal Article Reporting Standards (APA & Working Group on Journal Article Reporting Standards, 2008) has specified standards that include reporting confidence intervals along with effect sizes to elevate psychological and social science research transparency. Recommendations from entities such as the American Education Research Association (AERA; 2006) and the EQUATOR Network (Enhancing the QUAlity and Transparency of health Research; 2009) outline guidelines for reporting statistical results, and specifically call for measures of magnitude (i.e., effect size) and interquartile ranges (i.e., confidence intervals; AERA, 2010; Simera et al., 2010). However, the adoption of these best practice recommendations in reporting effect sizes along with NHSST results has been limited at best within various fields. For instance, empirical investigations into reporting practices in the fields

Table 2. Frequency and Percentage of Reporting Specific Effect Sizes for Six Journals That Publish Health Education and Behavior Research, 2010-2013.

Effect size type	*American Journal of Health Behavior* (*N* = 125), *n* (%)	*American Journal of Health Promotion* (*N* = 87), *n* (%)	*Health Education & Behavior* (*N* = 77), *n* (%)	*Health Education Research* (*N* = 109), *n* (%)	*Journal of American College Health* (*N* = 92), *n* (%)	*Journal of School Health* (*N* = 107), *n* (%)	Total[a]
Variance-accounted-for							
r/r^2	40 (32)	16 (18.3)	29 (37.6)	31 (28.4)	30 (28.0)	16 (17.3)	162
Eta-squared/ partial eta-squared (η^2)	8 (6.4)	9 (10.3)	2 (2.5)	9 (8.2)	16 (14.9)	2 (2.1)	46
R^2	6 (4.8)	2 (2.2)	10 (12.9)	8 (7.3)	5 (4.6)	4 (4.3)	35
Phi coefficient (φ)	2 (1.6)	0 (0)	0 (0)	0 (0)	4 (3.7)	1 (1.0)	7
Cramer's V (φ_c)	1 (0.8)	2 (2.2)	0 (0)	0 (0)	1 (0.9)	2 (2.1)	6
Spearman's rho	1 (0.8)	0 (0)	0 (0)	0 (0)	1 (0.9)	3 (3.2)	5
Omega-squared (ω^2)	1 (0.8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1
Rank biserial correlation (r_{rb})	0 (0)	0 (0)	1 (1.2)	0 (0)	0 (0)	0 (0)	1
Standardized differences							
Cohen's *d*	7 (5.6)	3 (3.4)	8 (10.3)	9 (8.2)	4 (3.7)	10 (10.8)	41
Glass's delta (Δ)	1 (0.8)	2 (2.2)	1 (1.2)	3 (2.7)	3 (2.8)	0 (0)	10
Hedge's *g*	0 (0)	1 (1.1)	0 (0)	1 (0.9)	0 (0)	1 (1.0)	3
Cohen's *F* (f^2)	0 (0)	1 (1.1)	0 (0)	0 (0)	0 (0)	1 (1.0)	2
Root mean square standardized effect (ψ)	1 (0.8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1
Cohen's *h*	0 (0)	1 (1.1)	0 (0)	0 (0)	0 (0)	0 (0)	1
Other measures							
Odds ratio/ adjusted odds ratio	77 (61.6)	51 (58.6)	34 (44.1)	62 (56.8)	41 (39.2)	75 (81.5)	340
Relative risk	2 (1.6)	6 (6.8)	2 (2.5)	0 (0)	2 (1.8)	0 (0)	12
Not stated[b]	0 (0)	0 (0)	1 (1.2)	0 (0)	0 (0)	1 (1.0)	2

Note. The unit of measurement is the number of individual effect size types reported. Several articles contributed to multiple effect size categories because they reported more than one effect size type in the published article. Thus, the total number of effect size types reported was *n* = 675, whereas the total number of article that reported an effect size type was *n* = 597.

[a]The total number of articles reviewed (*n* = 597). This column reflects the total number of effect sizes reported (*n* = 673). [b]Two articles failed to report the specific effect size employed and were thus excluded.

of psychology, counseling, special education, and general education suggest widespread poor effect size reporting practices (Thompson, 1999d). Kirk (2001) documented that in psychology-focused journals only 12% of published investigations included a measure of effect size. In examining five decades of published articles in the *Journal of Applied Psychology*, Finch et al. (2001) contended that there have been few changes in effect size reporting for half a century.

While the results reported here are by no means encouraging—as fewer than half of all examined articles reported an effect size of any type—it should be noted that journals with exceedingly high impact factors, and generally considered premier scientific outlets (i.e., *Nature* and *Science*), exhibit equally poor effect size reporting practices (Tressoldi, Giofre, Sella, & Cumming, 2013). Thus, many fields and several scientific outlets fare equally poorly as compared to the health education and behavior journals examined here.

Kirk (2001) noted that one potential reason underlying poor reporting practices of researchers is ignorance or misunderstanding of NHSST and associated *p* values. Vacha-Haase et al. (2000; Vacha-Haase & Thompson, 2004) contended that researchers fail to report effect sizes because

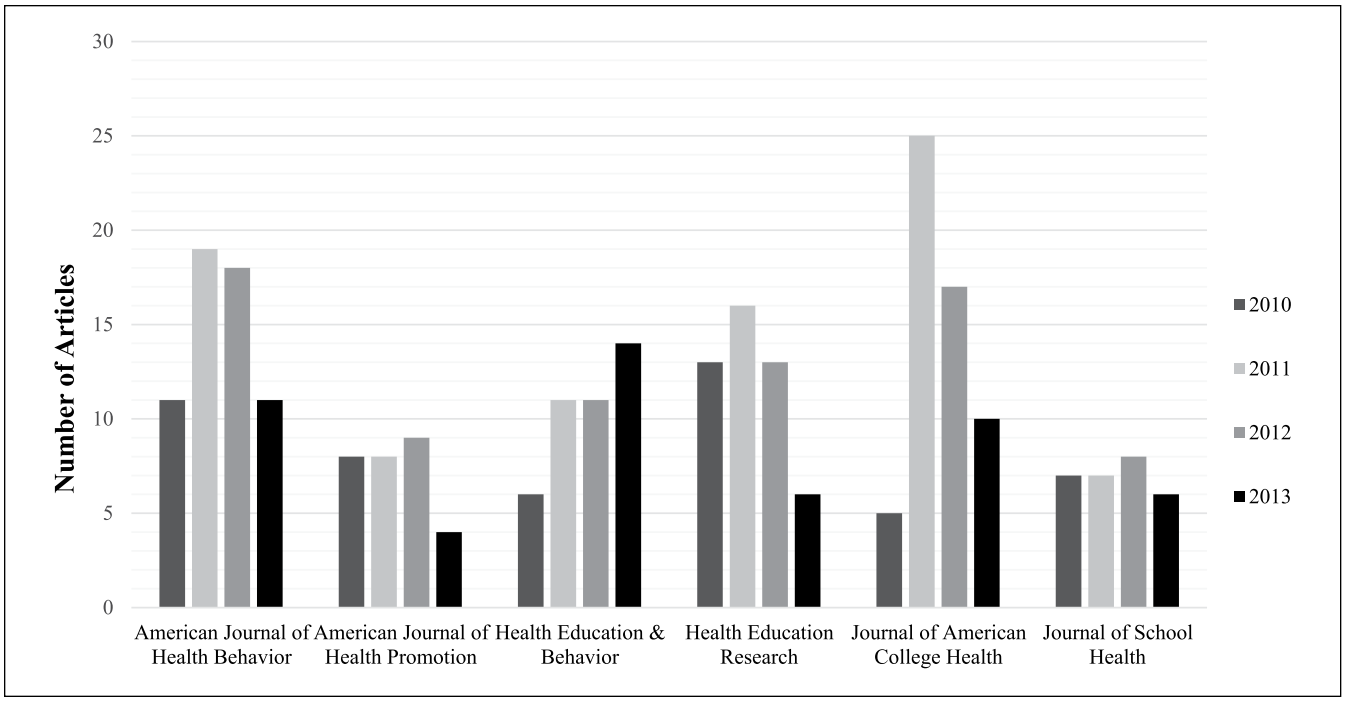

Figure 2. Comparison of variance-accounted-for effect size reporting for six journals that publish health education and behavior research, 2010-2013.
Note. These are absolute numbers of articles not adjusted by type of study.

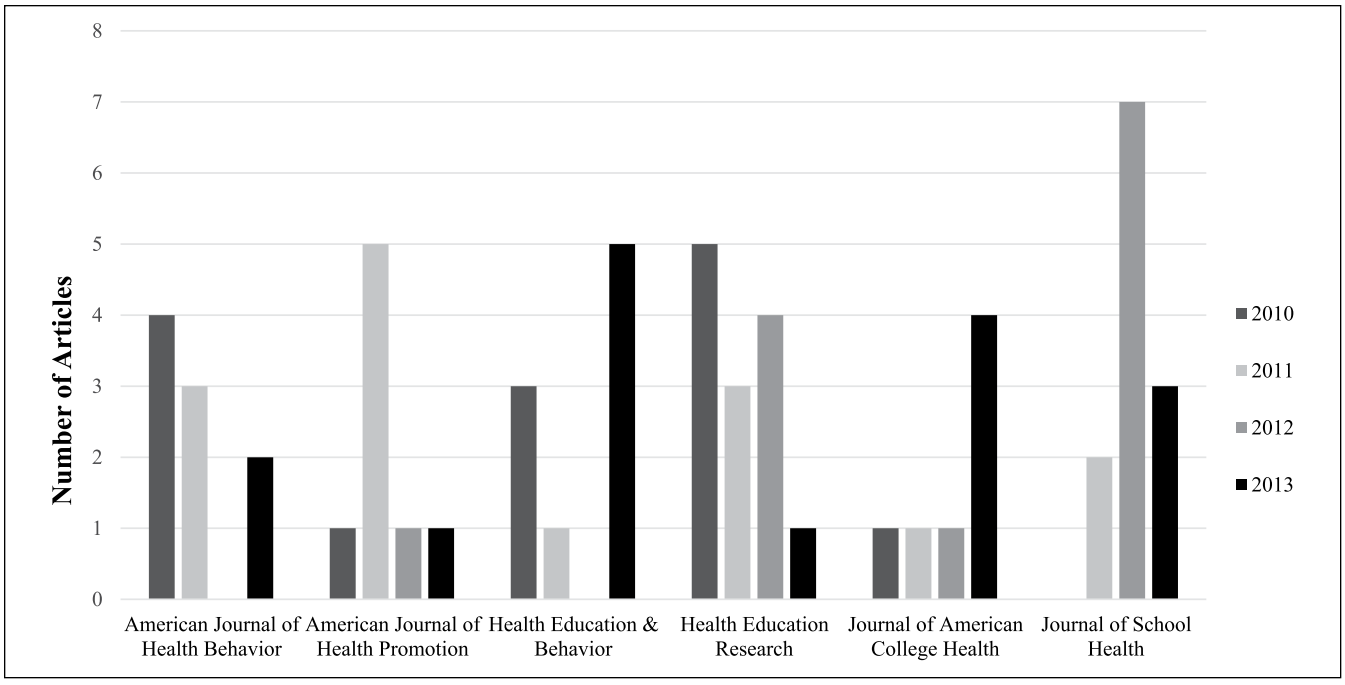

Figure 3. Comparison of standardized difference effect size reporting for six journals that publish health education and behavior research, 2010-2013.
Note. These are absolute numbers not adjusted by type of study.

(a) effect sizes were not emphasized in their training, and as a result (b) many researchers do not know how to compute and interpret effect sizes. Paul and Plucker (2004) asserted that effect size reporting is too infrequent because researchers

are unaware of the differences between statistical significance and effect size indices.

Our results suggest that health education and behavior researchers either do not interpret their effect sizes at all or rely heavily on the fixed conventions established by Cohen (1988). Thompson (2001) noted the problems associated with continued usage of small, medium, and large benchmarks, despite recommendations to evaluate effects within the particular context of the study, asserting, "If people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (pp. 82-83).

Effect Size Context Matters (Very) Much

As Thompson (2006b) has emphasized, effect sizes ought not to be interpreted against rigid benchmarks, because the research context matters so much. Even small effect sizes may be noteworthy in some contexts. So-called "small" effects may be noteworthy if

1. the outcome variable is very important, such as human longevity;
2. the outcome variable is especially resistant to intervention (Prentice & Miller, 1992);
3. small effects over time cumulate into large effects (Abelson, 1985); or
4. outcomes have multiple causes such that one or a few interventions inherently have limited impact (Ahadi & Diener, 1989; Strube, 1991).

For example, the eta squared for the effects of not smoking on longevity and of aspirin-taking as a prophylactic against heart attack are both only about 1%. But the outcome variable is so important! As Gage (1978) pointed out,

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

Limitations

There are several limitations of this analysis that should be noted. First, the research contained and being reported in the six journals selected for review are not representative of the entire body of health education and behavior research being conducted and published nationally and/or internationally. Thus, relying solely on data from the articles published in these specific journals limits our generalizability of statistical reporting practices across all domains of health research. While the peer-reviewed journals included in this study represent "flagship" journals for major health education professional organizations within health education and health-related behavioral and social science, it is important to note that these journals are all based within the United States. Consequently, our analysis excluded other English-language international journals that have health education and behavior as their primary scope. Further investigations, including a broader scope of health-related journals, are needed to fully understand statistical reporting practices of researchers.

Second, this investigation only evaluated reporting of effect size indices and did not include interquartile range (i.e., confidence intervals) reporting practices. Future studies should evaluate effect size and interquartile range reporting practices in accordance with the recommendations from numerous governing bodies on statistical best practices.

Finally, our results do not account for the type of study reported or the appropriateness of the reported effect size. Thus, a low rate of reporting a certain kind of effect size in a particular journal may be explained by that journal not having published any articles for which that type of effect size is appropriate.

Conclusion

As Fan (2001) contended, NHSST and effect size are two related sides of the same coin; they complement each other but cannot substitute for one another. Good research and good reporting practice require that both sides of the coin be taken into account to reach sound quantitative research decisions. Thus, for health education and behavior journals to be best positioned to positively influence health-related research, practice, and policy, it is essential that effect size reporting policies be explicitly outlined as a requirement for peer-reviewed publication. Guidelines from the AERA, APA, CONSORT (Consolidated Standards of Reporting Trials), and the EQUATOR Network provide a framework to enhance reporting practices. If the goal is to foster high-quality, transparent research, it is essential that editors of health education and behavior journals review their editorial policies in light of these and other guidelines and ensure that the articles they publish adhere to such reporting guidelines. Additionally, university faculty whose responsibility is the preparation of future health education and behavior researchers must carry the banner for effect size reporting and ensure that their students are not only familiar with the statistical concepts of effect sizes but also prepared to report and interpret effect sizes in their own future research.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129-133. doi:10.1037/0033-2909.97.1.129

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology*, *56*, 398-406. doi:10.1037/0022-3514.56.3.398

American Education Research Association. (2006). *Standards for reporting on empirical social science research in AERA publications*. Retrieved from http://ltc-ead.nutes.ufrj.br/constructore/objetos/AERA_Reporting%20Research.pdf

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

American Psychological Association and Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839-851.

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. doi:10.2307/3803199

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health*, *75*, 267-270. doi:10.1111/j.1746-1561.2005.tb06687.x

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*, 12-18. doi:10.1177/1090198113483139

Buhi, R. E. (2005). The insignificance of "significance" tests: Three recommendations for health education researchers. *American Journal of Health Education*, *36*, 109-112. doi:10.1080/19325037.2005.10608167

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292. doi:10.1080/00220973.1993.10806591

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003. doi:10.1037/0003-066X.49.12.997

Colliver, A. J. (2002). Call for greater emphasis on effect-size measures in published articles in teaching and learning in medicine. *Teaching and Learning in Medicine*, *14*, 206-210. doi:10.1207/S15328015TLM1404_1

Fan, X. T. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*, 275-282. doi:10.1080/00220670109598763

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology, and ecology. *Journal of Socio-Economics*, *33*, 615-630. doi:10.1016/j.socec.2004.09.035

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

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390. doi:10.1037/1082-989X.1.4.379

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

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*, 135-140. doi:10.1053/j.seminhematol.2008.04.003

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

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

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury.

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

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759. doi:10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213-218. doi:10.1177/00131640121971185

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *The American Psychologist*, *56*, 16-26. doi:10.1037//0003-066X.56.1.16

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in biomedical journals: The "statistical analyses and methods in the published literature" or the SAMPL guidelines." In P. Smart, H. Maisonneuve, & A. Polderman, (Eds.), *Science editors' handbook* (2nd ed., pp. 175-182). London, England: European Association of Science Editors.

Maher, J., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE: Life Sciences Education*, *12*, 345-351. doi:10.1187/cbe.13-04-0082

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834. doi:10.1016/j.appsy.2004.02.001

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*, 264-269. doi:10.1136/bmj.b2535

Nickerson, S. R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301. doi:10.1037/1082-989x.5.2.241

O'Fallon, J. R., Duby, S. D., Salsburg, D. S., Edmonson, J. H., Soffer, A., & Colton, T. (1978). Should there be statistical guidelines for medical research papers? *Biometrics*, *34*, 687-695.

Onwuegbuzie, A. J., Levin, J. R., & Leech, N. L. (2003). Do effect-size measures measure up? A brief assessment. *Learning Disabilities*, *1*, 37-40.

Paul, K. M., & Plucker, J. A. (2004). Two steps forward, one step back: Effect size reporting in gifted education research from 1995-2000. *Roeper Review*, *26*, 68-72. doi:10.1080/02783190409554244

Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review*, *20*, 122-126. doi:10.1080/02783199709553873

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160-164. doi:10.1037/0033-2909.112.1.160

Richardson, J. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*, 135-147. doi:10.1016/j.edurev.2010.12.001

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181. doi:10.1037/0003-066X.47.10.1173

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129. doi:10.1037/1082-989X.1.2.115

Simera, I., Moher, D., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). A catalogue of reporting guidelines for health research. *European Journal of Clinical Investigation*, *40*, 35-53.

Smith, M. L. (2009). Instrumentation and psychometric property reporting in current health education literature. *American Journal of Health Studies*, *24*, 232-239.

Strube, M. J. (1991). Multiple determinants and effect size: A more general method of discourse. *Journal of Personality and Social Psychology*, *61*, 1024-1027. doi:10.1037/0022-3514.61.6.1024

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434-438. doi:10.1002/j.1556-6676.1992.tb01631.x

Thompson, B. (1993). The use of statistical significance in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377. doi:10.1080/00220973.1993.10806596

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30. doi:10.2307/1176337

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them. *Theory & Psychology*, *9*, 165-181. doi:10.1177/095935439992006

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 & Psychology*, *9*, 191-196. doi:10.1177/095935439992007

Thompson, B. (1999d). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology*, *133*, 133-140. doi:10.1080/00223989909599728

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93. doi:10.1080/00220970109599499

Thompson, B. (2004). The "significance" crisis in psychology and education. *Journal of Socio-Economics*, *33*, 607-613. doi:10.1016/j.socec.2004.09.034

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

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583-603). Washington, DC: American Educational Research Association.

Thomson Reuters. (2016). *2015 Journal citation reports social science edition* (Public, Environmental and Occupational Health category). Retrieved from http://about.jcr.incites.thomsonreuters.com/full-titles-2015.pdf

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2.

Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *Journal of General Psychology*, *136*, 261-269. doi:10.3200/GENP.136.3.261-270

Tressoldi, P. E., Giofre, D., Sella, F., & Cumming, G. (2013). High impact = high statistical standards? Not necessarily so. *PLoS One*, *8*(2), e56180. doi:10.1371/journal.pone.0056180

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement*, *61*, 219-224. doi:10.1177/00131640121971194

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. doi:10.1177/0959354300103006

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*, 473-481. doi:10.1037/0022-0167.51.4.473

Westover, M. B., Westover, K. D., & Bianchi, M. T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine*, *9*, 1-20. doi:10.1186/1741-7015-9-20

Wilkinson, L., APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. doi:10.1037/0003-066X.54.8.594