Should We Be Confident in Published Research? A Case Study of Confidence Interval Reporting in Health Education and Behavior Research

Barry, A.E, PhD*
Reyes, J.V., PhD, CHES
Szucs, L.E., PhD, CHES
Goodson, P., PhD
Valdez, D., PhD

Abstract

Confidence intervals (CIs) have been highlighted as "the best" reporting device when reporting statistical findings. However, researchers often fail to maximize the utility of CIs in research. We seek to (a) present a primer on CIs; (b) outline reporting practices of health researchers; and (c) discuss implications for statistical best practice reporting in social science research.

Approximately 1,950 peer-reviewed articles were examined from six health education, promotion, and behavior journals. We recorded: (a) whether the author(s) reported a CI; (b) whether the author(s) reported a CI estimate width, either numerical or visual; and (c) whether an associated effect size was reported alongside the CI.

Of the 1,245 quantitative articles in the final sample, 46.5% (n = 580) reported confidence interval use; , and 518 provided numerical/visual interval estimates. Of the articles reporting CIs, 383 (64.2%) articles reported a CI with an associated effect size, meeting the American Psychological Association's (APA) recommendation for statistical reporting best-practice.

Health education literature demonstrates inconsistent statistical reporting practices, and falls short in employing best practices and consistently outlining the minimum expectations recommended by APA. In an effort to maximize utility and implications of health education, promotion, and behavior research, future investigations should provide comprehensive information regarding research findings.

*Corresponding author can be reached at: aebarry@tamu.edu

Purpose and Rationale

Historically, when reporting data analyses and findings most health behavior researchers and other social scientists have relied almost exclusively on null hypothesis statistical significance testing (NHST) and p values (APA, 2010; Westover et al., 2011). These researchers either fail to recognize, choose to ignore, or (worse) may not be aware that statistical significance analysis constitutes merely a preliminary test requiring further contextualizing for valid interpretation, using additional information such as effect sizes and confidence intervals (CIs).

In a previous issue of *Health Education & Behavior* (Barry et al., 2016), an assessment of effect size reporting in manuscripts published within top-tier health promotion and health behavior journals is reported. This investigation, however, did not include an assessment of CI reporting. To address this oversight and further contextualize those previous findings, herein we report our assessment of the same dataset, specifically capturing whether researchers documented confidence interval estimates—either numerically or visually—and if an associated effect size was described alongside the confidence interval.

The focus on CIs is motivated by the numerous calls from various professional bodies. For instance, the Task Force on Statistical Inference (established in 1996 to analyze strategies and practices in statistical reporting) recommends researchers *always* provide CIs to help readers understand the quality of point estimates (APA, 2010; Thompson, 2007; Wilkinson & the APA Task Force on Statistical Inference, 1999). Aligned with this call, the Consolidated Standards of Reporting Trials (CONSORT)[1] Statement (developed in 1996, revised in 2001, 2007, and published in 2010) specifically addresses CIs as a metric researchers should provide in reports of all randomized control trials: "For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)" should be reported [item 17a in checklist]. Finally, the American Education Research Association (AERA, 2006), the EQUATOR Network [Enhancing the QUALity and Transparency of Health Research] (Lang & Altman, 2013), and the American Psychological Association (2008) have also put forth specific recommendations and guidelines for reporting CIs as a way to enhance psychological and social science research transparency.

Reporting Recommendations

Specified recommendations from entities such as the American Education Research Association [AERA] and the EQUATOR Network [Enhancing the QUALity and Transparency of health Research] (Lang & Altman, 2013) provide guidelines for reporting statistical results inclusive of magnitude of results (i.e., effect size) and interquartile ranges (i.e., CIs) as critical to health research. However, the adoption of these best practice recommendations in reporting CIs along with NHST results has been limited at best within various fields. The American Psychological Association (2008) not only "encouraged" (p. 18) the reporting of effect sizes, but further asserts researchers should "*always* [emphasis added] provide some effect size estimate when reporting a *p* value" (APA, 2008, p. 599). Moreover, recent recommendations from the APA's Publications and Communications Board and Working Group on Journal Article Reporting Standards (JARS) specified revised standards which include reporting CIs along with effect sizes to elevate psychological and social science research transparency (APA JARS Group, 2008).

These recommendations result from decades of debate and criticism concerning statistical reporting practices in the sciences and social sciences. Poor statistical reporting has been diagnosed as detrimental to the growth and refinement of scientific knowledge (Cumming & Fidler, 2009; Thompson, 2002) and, currently, all breeds of science find themselves steeped in controversies surrounding the validity of reported research, as well as their utility for replication attempts (Ioannidis, 2005).

The need for documenting and utilizing CIs is entangled in the on-going debate surrounding NHST and *p* value statements. For instance, in an editorial in the *Journal of American College Health*, an argument is offered for the insufficiency of *p*-value reporting which advocated for the inclusion of effect size measures to better ground and contextualize research findings—especially

[1] The CONSORT Statement emerged from the work of two groups comprised of scientists, medical journal editors, epidemiologists, and methodologists in Canada and the United States. Concerned with the absence of critical information in research reports, the groups combined their efforts and produced a checklist of elements that should be reported in any and all accounts of randomized trials. (see http://www.consort-statement.org/about-consort accessed Jan 09. 2019).

non-significant ones ($p > .05$, for example) (Barry et al., 2019). Simply put, NHST provides information on how likely the occurrence is of a resulting statistic, for the particular sample being examined. If the likelihood of a given result is small (commonly used threshold is $< .05$), researchers tend to conclude 'something is going on' with the sample being studied—the resulting statistic has a small probability of having occurred merely by chance. But the statistic obtained during analysis is never absolutely precise, even if the probability of obtaining it by chance is small: a given amount of error is always present and the result from a specific sample may, indeed, be an error. The questions facing researchers are always: "How much error surrounds my findings/results?" and, "How willing am I to be transparent about the errors in my study?"

CIs help answer the first question; researchers' sound reporting habits answer the second one. CIs provide the context for understanding a parameter estimate – a context that takes into account the amount of error present in the estimation. CIs, therefore, provide a layer of precision and 'reality-check' for the estimated parameter. Much in the same way researchers always report a standard deviation (SD) — or the 'spread' of scores — when describing an average/mean value, CIs provide a sense of how much error is 'spread around' the resulting statistic, to stretch the analogy. CIs, therefore, provide information about the precision and reliability of the parameters being estimated (Belia et al., 2005; Cumming & Finch, 2001).

Another important use of CIs is in meta-analyses, as they allow for comparisons among similar studies. Such comparisons are the basis for the "acquisition of cumulative knowledge" (Hubbard, 2015, p. 70). When CIs in different studies of the same phenomenon overlap, the overlap suggests "credible estimates of the same population parameter(s)" (Hubbard, 2015, p. 70). In his argument proffering the notion of "significant sameness" as opposed to "significant difference," Raymond Hubbard details several advantages of CI reporting for building cumulative knowledge in a given field. Among these advantages, he describes how CIs are able to indicate whether a replication was successful; how CIs are able to "sidestep the baneful effects of low statistical power common in traditional significance testing" (p. 76); and how CIs help highlight "commonalities in data sets, the road to generalization" (p. 75) Hubbard also points out that analyses of CIs can help prevent over-reliance on single studies when building knowledge in a given field (Hubbard, 2015). One final advantage of CIs is their ease of reporting. They can be depicted simply in a visually-intuitive manner, using a horizontal line with end-point anchors (Thompson, 2002).

Given the advantages and the contributions CIs make to research and knowledge building, one is left to wonder, along with Hubbard, "…why CIs—a procedure that Tukey (1960) viewed as probably the most important among all types of statistical methods we know—are not routinely used, reported, and interpreted" (Hubbard, 2015, p. 70). Even though in this paper we do not answer the question about *reasons* for low CI reporting (this would require an entirely different study), we outline health education, promotion, and behavior researchers' common practices with the intent of highlighting strengths and weaknesses prevalent in professional publications, and call for their improvement. In particular, we seek to establish the extent to which published health promotion and behavior research meets the established "best practices" in statistical reporting – reporting CIs along with an effect size.

Methods

The methods employed herein mirror those described in detail within Barry et al. (2016). Briefly, the current review included: (a) examining four years of CI reporting practices among six journals in the field of health education, health promotion, and health behavior — *American Journal of Health Behavior (AJHB), American Journal of Health Promotion (AJHP), Health Education & Behavior (HEB), Health Education Research (HER), Journal of American College Health (JACH), and Journal of School Health (JoSH).* These journals were specifically examined because they represent the premiere journals associated with national health education and health behavior professional societies, and have been the focus of previous investigations examining statistical reporting in the field of health behavior and promotion (Barry, 2005; Barry et al., 2014); (b) reviewing 1,950 refereed articles published between 2010 and 2013, the same time-frame and sample reported on by Barry et al. (2016); and (c) in addition to documenting relevant bibliographical information (authors, year of publication, journal source, volume, pages), assessing whether the author(s) reported a confidence interval within the text of the article, reported a confidence interval estimate width – either numerical or visually, and/or described an associated effect size measure, alongside the confidence interval.

After excluding non data-based articles, such as commentaries (n = 70) and other published work that did not report or include quantitative data, such as literature reviews and qualitative studies (n = 545), a total of 1,245 published articles constituted the final sample. These investigations represent a comprehensive portrait of the health education, promotion, and behavior literature, spanning a total of 24 volumes across the six selected journals.

Results

Reporting of Cis

Table 1 presents the number of articles examined from each journal, including the annual distribution (in percentages) of CI reporting, across four years. Of the 1,245 articles in the final sample, 46.5% (n = 580) reported a CI. The annual reporting percentages ranged from a low of 27.6% (*JACH*, in 2011) to a high of 65.2% (*JACH*, in 2013). Figure 3 visually illustrates journals' CI reporting trends across the four year time period.

Of the total number of articles reporting CIs, n = 518 (89.3%) provided interval estimate widths, either numerically or visually. Across specific journals, numerical/visual interval estimate reporting ranged from 85.3% (*JACH)* to 91.1% (*AJHB)*. Of the 580 articles reporting CIs, 383 (64.2%) reported CIs as well as effect sizes, meeting the APA's recommendation for statistical reporting best practices. Table 1 presents the frequency of articles meeting the APA's recommendation by journal and year. Across all journals in the four-year period, the percentage of articles which demonstrated APA-recommended best practices reporting (providing CIs and effect sizes) ranged from 59.2% (*HEB)* to 69.1% (*AJHB*). These percentages (i.e., 69.1% *AJHB*) were calculated by dividing the total frequency of articles demonstrating APA recommended best practices for each journal (i.e., n = 85 *AJHB*) by the total frequency of articles (i.e., n = 123 *AJHB*) that reported CIs in the respective journal. Figure 4 illustrates the best practice publishing trends of paired confidence interval *and* effect size measures across the four year time period or each investigated journal.

Table 1

Annual Overall Reporting of Confidence Interval and APA best practice by Journal and Year

Journal Title and Years	Annual CI Reporting %	Frequency of CI Reporting			APA best practices in statistical reporting (*n*)
		Not Reporting (n)	*Reporting (n)*	*Total (n)*	
American Journal of Health Behavior		124	123	247	85
▪ 2010	50.8%	31	32	63	20
▪ 2011	59.4%	26	38	64	31
▪ 2012	41.8%	39	28	67	19
▪ 2013	47.2%	28	25	53	15
American Journal of Health Promotion		112	84	196	55
▪ 2010	42.9%	24	18	42	12
▪ 2011	47.1%	36	32	68	25
▪ 2012	40.4%	28	19	47	10
▪ 2013	38.5%	24	15	39	8
Health Education & Behavior		81	76	157	45
▪ 2010	42.1%	22	16	38	9
▪ 2011	34.3%	23	12	35	7
▪ 2012	60.0%	16	24	40	12
▪ 2013	54.5%	20	24	44	17
Health Education Research		99	115	214	74
▪ 2010	55.4%	29	36	65	24
▪ 2011	53.4%	27	31	58	19
▪ 2012	58.3%	20	28	48	19
▪ 2013	46.5%	23	20	43	12
Journal of American College Health		121	67	187	45
▪ 2010	30.0%	21	9	30	4
▪ 2011	27.6%	56	21	76	15
▪ 2012	37.9%	36	22	58	15
▪ 2013	65.2%	8	15	23	11
Journal of School Health		128	115	244	79
▪ 2010	37.0%	34	20	54	13
▪ 2011	45.6%	37	31	68	20
▪ 2012	55.8%	23	29	52	19
▪ 2013	50.7%	34	35	69	27
TOTAL ARTICLES SAMPLE	**46.6%**	**665**	**580**	**1245**	**383**

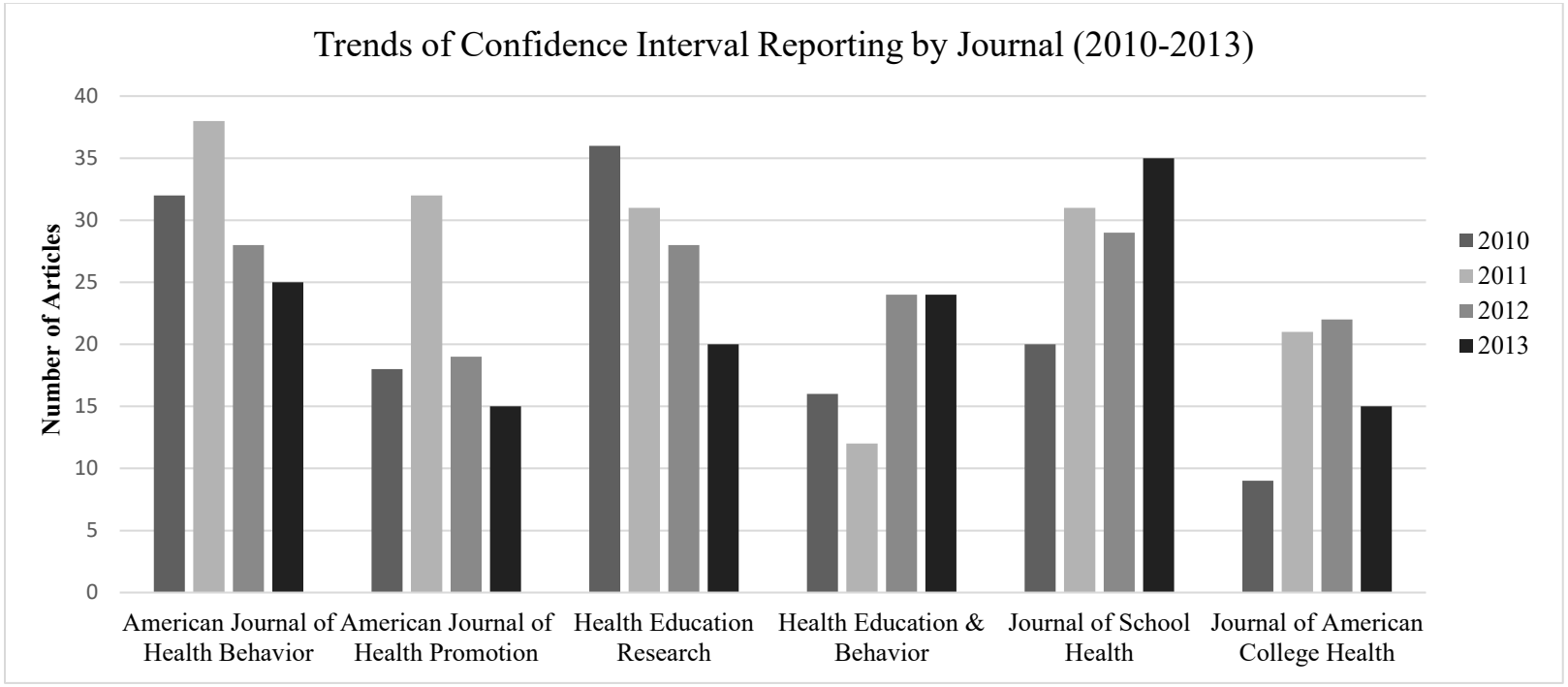

Figure 1. Comparison of CI reporting for six journals that publish health education and behavior research, 2010-2013. *Note.* These are absolute numbers of articles not adjusted by type of study.

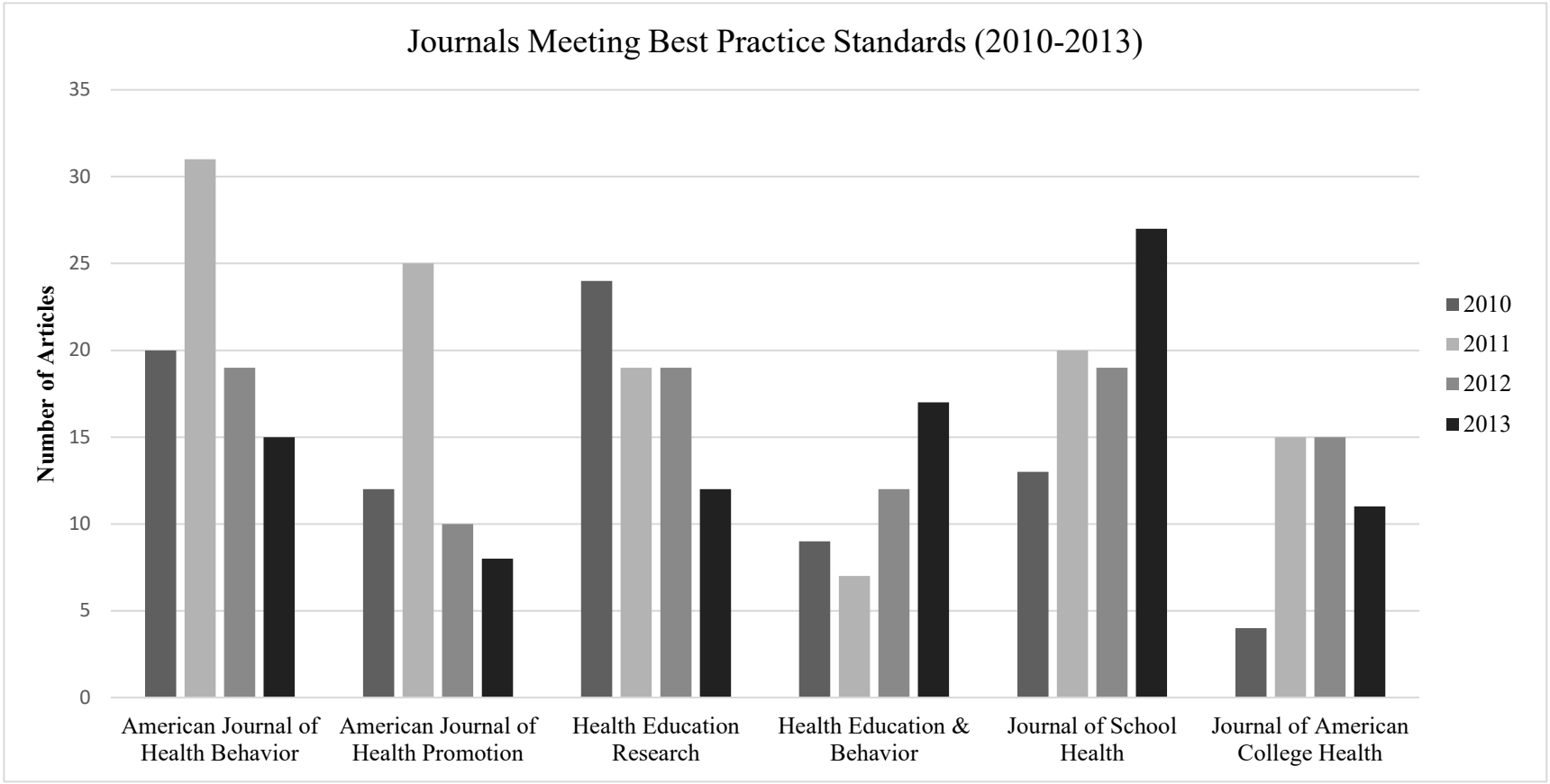

Figure 2. Comparison of best practice reporting for six journals that publish health education and behavior research, 2010-2013. *Note.* These are absolute numbers of articles not adjusted by type of study.

Discussion

Our review indicated less than half of the published literature we examined from six prestigious journals publishing health education, promotion, and behavior research—over the span of four years—adhered to recommendations and calls for reporting CIs. While at face value this scenario seems less than encouraging, when placed within the broader publishing landscape, researchers in health education, promotion, and behavior are reporting CIs more frequently than those in other applied sciences. For example, across 245 articles in 49 journal volumes of *Education Administration Quarterly* (Byrd, 2007), the majority of quantitative studies included effect sizes; however, no quantitative studies examined included CIs. In expanding their investigation of the educational administration field to 473 articles encompassing 95 volumes of two different journals, there was no CI reporting for any quantitative research (Byrd & Eddy, 2009). While our intent was to examine quantitative research in premiere journals associated with national health education and health behavior professional societies, it is important to note that health behavior research is published in scholarly outlets other than the six journals included in this investigation. Thus, more expansive and inclusive explorations of statistical reporting practices are warranted.

Although our review does not answer the questions regarding "reasons why," the sub-optimal reporting practices might be driven by lack of understanding of the purposes of CIs, misconceptions about their interpretive values, and/or misunderstandings about how intervals can enhance NHST (Schmidt, 1996; Schmidt & Hunter, 1997). Our hope is the primer presented above can serve to eliminate misconceptions among health education, promotion, and behavior researchers. The findings of the present investigation suggest that some researchers engage in the healthy practice of following the practical recommendations of the APA, by including both CIs and effect sizes. Unfortunately, a much larger percentage of researchers in health promotion and behavior are not. Our results mirror previous investigations documenting poor reporting of attrition (Barry, 2005) and validity/reliability characteristics (Barry et al., 2014). Additional exploration of the statistical reporting practices of health education, promotion, and behavior researchers is warranted, however, as continued discussion of these issues is paramount to the growth, sustainability, and implications of the broader field.

In order for health promotion research outlets to be best positioned to positively influence health-related research, practice, and policy, it is important that confidence interval reporting be outlined as a firm editorial recommendation for peer-reviewed publication. As of the writing of this article, however, none of the journals reviewed contained practical guidelines recommending that CI estimates should be included as a criteria for peer-review and publication. To keep this line of dialogue moving forward, professional preparation programs for future health education, promotion, and behavior researchers must be proactive to ensure their students are not only familiar with the significance of CI reporting, but demonstrate self-efficacy to report and interpret CIs in their own future research. In effect, if students and emerging professionals understand the merits of understanding and reporting CIs and effect sizes, there exists potential.

Regardless, the information highlighted herein intends to educate health education, promotion, and behavior researchers, and ultimately influence future work appearing in scholarly journals. By educating and advocating for inclusive reporting practices

via CIs and effect sizes, the field stands to gain from more thorough, high quality research to advance health education, promotion, and behavior forward. In doing so, the field may directly influence health policy and practices in the United States.

Conclusions

The 6th edition of the APA Publication Manual states that anytime a table includes point estimates, it should include the CIs (APA, 2010, p. 138). With more journals and research fields requiring documentation of effect sizes and CIs, we should push to encourage researchers to begin substantiating their hard-earned research findings with these measures of quality. Recognizing that encouragement alone is likely insufficient, we contend that scholarly publication outlets develop guidelines and requirements outlining how submissions to the journal must align with APA recommendations and statistical reporting best practices. Such policies are currently missing from the journals included in this review. By shifting requirements and paradigms in statistical method reporting, we will begin to more fully understand the health phenomena occurring in society. Incorporating both CIs and effect sizes will help readers better grasp and contextualize the results of the study, advance the field, and ultimately position practitioners to better influence health behaviors.

Acknowledgments

The authors have no conflict of interest to report, financial or otherwise.

References

American Education Research Association (AERA). (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40. https://doi.org/10.3102/0013189X035006033

American Psychological Association. (2008). *Publication manual of the American Psychological Association* (4th ed.). American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). American Psychological Association.

American Psychological Association (APA) Working Group on Journal Article Reporting Standards (JARS Group) (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9): 839-851. https://doi.org/10.1037/0003-066X.63.9.839

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health, 75(7),* 267-270. https://doi.org/10.1111/j.1746-1561.2005.tb06687.x

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Educ. Behav, 41*(1), 12-18. https://doi.org/10.1177/1090198113483139

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes: The handling

of quantitative results in published health education and behavior research. *Health Education & Behavior, 43*(5), 518-527. https://doi.org/10.1177/109019811666952 1

Barry, A. E., Valdez, D., Goodson, P., Szucs, L. E., & Reyes, J. V. (2019). Moving college health research forward: Reconsidering our reliance on statistical significance testing. *Journal of American College Health, 67*(3), 1-8. https://doi.org/10.1080/07448481.2018.1 470091

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389-396.

Byrd, J. K. (2007). A call for statistical reform in EAQ. *Educational Administration Quarterly, 43*(4), 381-391. https://doi.org/10.1177/0013161X062971 37

Byrd, J., & Eddy, C. (2009). Statistical applications in two leading educational administration journals. *Journal of Educational Administration, 47*(4), 508-520. http://dx.doi.org/10.1108/095782309109 67473

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Journal of Psychology, 217*(1), 15-26. https://doi.org/10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532-572. https://doi.org/10.1177/00131644016140 02

Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. Sage Publications.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.002 0124

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*. European Association of Science Editors. http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129. https://doi.org/10.1037/1082-989X.1.2.115

Schmidt F., & Hunter J. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-63). Erlbaum.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32. https://doi.org/10.3102/0013189X031003025

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423-432. https://doi.org/10.1002/pits.20234

Tukey, J. W. (1960). Conclusions vs. decisions. *Technometrics, 2*(4), 423-433.

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

Wilkinson L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594-604. https://doi.org/10.1037/0003-066X.54.8