A Closer Look at Effect Sizes and Their Relevance to Health Education

Daphne C. Watkins, Desiree Rivers, Kyrel L. Rowell, B. Lee Green, and Brian Rivers

Abstract

A number of academic disciplines are engaged in scholarly discussions regarding statistical practice reform, particularly the use of effect sizes. Health education must stimulate a similar conversation by adopting strategies for generating, reporting, and interpreting effect size estimates for various statistical analyses within journal articles. The purpose of this article is to demonstrate the practical applications of effect size reporting and interpretation in health education research. Congruent with previous recommendations in the American Journal of Health Education, this article will provide examples and techniques used for effect size reporting that educate the researcher and practitioner, thus improving the scholarship of health education publications. Effect size reporting should become the rule for health education and concerted efforts should be made to equip researchers and practitioners with the proficiency to perform this task effectively. Such skill building will increase the scholarship and readability of health education research.

For more than a decade, the scientific community has grappled with the most effective way to report statistical information.[1] Merrill, Stoddard, and Shields[2] summarized how statistical procedures are used by health educators and *acknowledged* the need for an increase in the use of statistical techniques by these professionals. Subsequently, Buhi[3] suggested three practices that health educators should use to supplement statistical significance testing, and *advocated* the need for calculating effect sizes. The purpose of this article is to demonstrate the practical applications of effect size reporting and interpretation in health education research. This will be achieved by (1) defining effect sizes; (2) identifying the three classes of effect sizes; (3) providing practical applications for reporting and interpreting effect sizes; and (4) suggesting guidelines to demonstrate applications that will improve the scholarship of health education journals. Specifically, this article will examine effect size reporting and interpretation with commonly used statistical procedures in public health:[4] analysis of variance (ANOVA) and regression.

EFFECT SIZES

The importance of reporting effect sizes was first introduced in the mid-1970s.[5] Throughout the proceeding decades, however, researchers continued to support the utilization of *p* values as the primary indicator of statistical significance. Although statistical significance "evaluates the probability or likelihood of the sample results, given the sample size, and assuming that the sample came from a population in which the null hypothesis is exactly true," it does not evaluate the *importance* of the results.[6]

Daphne C. Watkins is a doctoral candidate in the Division of Health Education and a graduate research assistant for the Center for the Study of Health Disparities at Texas A&M University, 4243 TAMU, College Station, Texas 77843; E-mail: dwatkins@tamu.edu. Desiree Rivers, MSPH, Kyrel L. Rowell, MPH, B. Lee Green, PhD, and Brian Rivers, PhD, are all with the Center for the Study of Health Disparities at Texas A&M University, 4243 TAMU, College Station, Texas 77843.

Conversely, estimating and interpreting effect sizes do provide the researcher with such a judgment regarding the *practical significance* of study results.[7] Additionally, effect sizes describe the significance of findings in relationship to practice and present findings in ways that are readily understood by professionals in the field.[8] When they are reported and interpreted effectively, effect sizes are practical, straightforward, and relevant to research questions and hypotheses.[8]

Over sixty different standardized statistics are used for analyses;[9] however, the most widely used effect size statistics in the behavioral sciences are Cohen's *d*, Glass's delta (Δ), eta^2, adjusted R^2, and omega squared. These more popular effect size estimates will be referenced throughout this article. Effect sizes not only apply to univariate means comparison, but also to multiple regression (or prediction), multiple correlation, multivariate analysis of variance (MANOVA), and univariate proportion comparisons.[10] A review of textbooks that considered the use of statistical significance and effect size suggested that statistical testing will continue; however, effect size reporting will become prevalent as journal editors revise editorial policies to reflect recommendations of the APA *Publication Manual.* [11] Specifically, the fifth edition of the APA *Publication Manual* along with the Task Force on Statistical Significance (appointed in 1996) called for the reporting of effect sizes. The Task Force argued that:

> It is almost always necessary to include some index of effect size or strength of relationship...the general principle to be followed...is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. [12]

Wilkinson and the APA Task Force on Statistical Inference further assert that researchers should always provide some effect-size estimate when reporting a *p* value.[12]

Although estimating effect sizes is strongly encouraged by APA, it has been noted that merely 'encouraging' effect size reporting will send a self-canceling mixed-message to authors.[12] To 'encourage' authors to report effect sizes while 'requiring' them to abide by the strict rules of author note placement, pagination, and margins is sending the message "these myriad requirements count; this encouragement doesn't."[13] To highlight the importance of effect sizes, emphasis should be placed on requiring authors to report and interpret effect sizes in health education journals.

CLASSES OF EFFECT SIZES

Vacha-Haase and Thompson[14] identify three major classes of effect sizes. These classes are further illustrated in Table 1. *Variance-accounted-for* statistics are highlighted in this article.

Effect Size Procedures

Statistical Package for the Social Sciences (SPSS v.11.0) is commonly used by researchers of health education and will be used in this article to describe analyses techniques. Despite the statistical package preferences used in this article, other packages (i.e., SAS, STAT, etc.) can perform similar techniques and produce the same results. SPSS does not yield standardized difference effect sizes (such as Cohen's *d* or Glass's Δ); however, the computations are fairly simple and can be created using a calculator or a spreadsheet.[15] Table 2 offers a detailed explanation of some common analytic methods and illustrates how effect sizes can be used to acquire the results in SPSS. Moreover, in this article the two techniques that will be discussed are ANOVA and regression.

Effect Size in Analysis of Variance (ANOVA) Procedures

Eta^2 (or, $\eta 2$) is an estimate of the proportion of variability in the dependent variable(s) explained, or accounted for; by

Table 1. Three Major Classes of Effect Sizes

Standardized differences effect sizes	▪ In order to do an "apples-to-apples" comparison of effects across the literature where researchers have used different methods, standardized differences must be reported. ▪ Standardized differences articulate effect sizes in standard deviation units.
Variance-accounted-for effect sizes	▪ Variance-accounted-for effect sizes can be computed analogously for multivariate analyses, given the general linear model (GLM). ▪ Variance-accounted-for effect sizes for such data are difficult to interpret with respect to practical significance. As a result, for categorical data, effect sizes (i.e. the binomial effect size) display or odds ratios are recommended.
"Corrected" effect sizes	▪ Corrected effect sizes can occur if we can successfully estimate the amount of sampling error variance in the sample data and then remove this influence from the effect size. ▪ Corrected effect sizes try to improve the estimate of either the population or future sample effects by removing the estimated influences of sample peculiarity from the results.

Note: The only time the sample effect size will not be inflated by sampling error is if the sample is representative of the entire population (or the population effect size is perfect). Sampling error is affected by sample size, the number of measured variables, and the population effect size.

Table 2. Strategies for obtaining effect sizes for selected SPSS Analyses

Analysis	Possible strategy
Contingency table (r or odds ratio)	▪ Run the CROSSTABS procedure and select the desired effect from the STATISTICS submenu.
Independent t test (d, η^2, ω^2)	▪ Compute a Cohen's d by hand. Or, run the analysis as a one-way ANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. Use the Hay's correction formula (ω^2) if an adjusted estimate is desired.
ANOVA (η^2 or ω^2)	▪ Run the analysis as an ANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. Use Hay's correction formula by hand if an adjusted estimate is desired.
Regression ($R\_$ or R_*)	▪ Run the REGRESSION procedure. Both the uncorrected R^2 and the corrected variance accounted for (R^2*) estimates are displayed, by default.
MANOVA (multivariate η^2 or ω^2)	▪ Run the analysis as a MANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. A corrected estimate, multivariate ω^2, can be computed by hand.
Descriptive discriminant analysis (multivariate η^2 or ω^2)	▪ Run the analysis as a MANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. A corrected estimate, multivariate ω^2, can be computed by hand.
Canonical correlation analysis ($R\varsigma^2$ or $R\varsigma^{2}*$)	▪ Run the analysis in the MANOVA procedure using the syntax suggested by Thompson.[15] The R^2 is reported. Apply the Ezekiel correction by hand if a corrected value ($R\varsigma^{2}*$) is desired.

Note. ANOVA = analysis of variance; GLM = general linear model; MANOVA = multivariate analysis of variance

Source: Vacha-Haase & Thompson, 2004.

membership in the groups defining the independent variable.[8] SPSS provides estimates of effect size in the form of eta^2 and the researcher must specify that the effect size estimate is preferred through the commands. Another name for eta^2 estimates is *variance-accounted-for* statistics, as previously mentioned. Under these estimates (1) all parametric analyses are correlational, (2) all such analyses apply weights (i.e., a, b, and β weights), and (3) the latent variable (or underlying effect) scores are the focus of all analyses.[16]

Eta^2 is achieved in SPSS via the (1) *Analyze*, (2) *Compare Means*, and (3) *Means* procedures; or through the (1) *Analyze*, (2) *General Linear Model*, (3) *Univariate, Multivariate*, or *Repeated Measures* procedures. Consider Figure 1, where eta^2 and partial eta^2 are presented. Researchers should note that when reporting the effect size in ANOVA, the printout will provide the overall effect size estimate by default. However, in order to obtain the effect sizes of individual factors, or the partial eta^2 for each individual factor, the researcher must specify this request via the (1) *Analyze*, (2) *General Linear Model*, and (3) *Univariate, Multivariate*, or *Repeated Measures* procedures from the drop-down menu in SPSS. Readers are cautioned that the partial eta^2 for each individual factor does not add up to the total explained variance. Additionally, when the subject-to-variable ratio is small, the explained variance is biased and may produce misleading results. In lieu of this, researchers should consider reporting the adjusted effect size (or the adjusted eta^2), which will provide an unbiased result. One would interpret the eta^2 value in Figure 1 as, "regarding effect size, about 70% of the variability in the dependent variable scores was explained with the knowledge of group membership on the independent variable."

Effect Size in Regression Procedures

Effect sizes are commonly reported in regression analyses in the form of R^2, another *variance-accounted-for* effect size. SPSS reports both the uncorrected and the corrected R^2 in regression analyses. Interpretations of effect sizes in regression analyses can be illustrated using an R^2 of .32. The interpretation of this statistic would be "regarding effect size, 32% of the variability in the dependent variable was explained or accounted for, by the independent variables." The researcher should note that the effect size for each independent variable can be accomplished by calculating an eta^2 change in stepwise to determine the effect size of each predictor. Due to the ongoing debate surrounding the pros and cons of using stepwise, readers are encouraged to view the work of Thompson[17] and others who present advantages and disadvantages of stepwise as a statistical technique.

Figure 1. Example of Eta^2 and Partial Eta^2

ANOVA Table

			Sums of Squares	df	Mean Square	F	Sig.
DV * LEVEL	Between Groups	(Combined)	722.667	5	144.533	2.891	.114
	Within Groups		300.000	6	50.000		
	Total		1022.667	11			

Measures of Association

	Eta	Eta^2
DV * LEVEL	.841	.707

Tests of Between-Subjects Effects (Dependent Variable: DV)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	**Partial Eta^2**
Corrected Model	722.667	5	144.533	2.891	.114	.707
Intercept	4181.333	1	4181.333	83.627	.000	.933
LEVEL	722.667	5	144.533	2.891	.114	.707
Error	300.000	6	50.000			
Total	5204.000	12				
Corrected Total	1022.667	11				

R Squared = .707 (Adjusted R Squared = .462)

Interpreting the magnitude of Effect Sizes

Kline[18] noted that when researchers first learn about effect sizes they ask questions such as: *What is a large effect? What is a small effect? What is a substantive (important) effect?* The magnitude (or size) of the change is important with effect sizes and not as valued with statistical significance testing. With the magnitude of effect sizes in mind, Jacob Cohen[19] devised guidelines to address the first two questions. The third question will be addressed later in this section. This article emphasizes the importance of interpreting effect sizes in relation to prior studies and not using Cohen's benchmarks for what may be considered 'small,' 'medium,' and 'large' effects. According to Thompson, "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."[20] Kline[18] on the other hand, supported the benchmarks but identified seven cautions about interpreting these guidelines set by Cohen. Readers are encouraged to view the work of Kline[18] for more information regarding these cautions.

Answering the question "*What is a substantive effect?*" is a difficult task. According to Kline, [18] deciding whether an effect is important or not is complicated because expressing an effect's significance (i.e., theoretical, practical, or clinical) requires more discipline-specific expertise than it does when estimating its magnitude. Taking this into consideration, the answer depends on the research context. Each effect size reported is strongly related to the research question under investigation, such that a large effect may have as much substantive significance as a small effect. Using a smoking cessation example with a small effect, Gage noted:

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

In most studies (i.e., intervention studies) a non-zero effect size is desired for the primary hypotheses. For example, if a health educator has designed a new smoking cessation program, his or her expectation is that the effect sizes will illustrate that the new program is more effective than traditional smoking cessation programs. On the other hand, if a health educator is reviewing the unexpected consequences, or perhaps side effects, he or she will want near-zero effect sizes.

REPORTING GUIDELINES

Vacha-Haase and Thompson,[14] in compliance with Wilkinson and the APA Task Force on Statistical Significance, suggested three guidelines for reporting effect sizes. These guidelines should be applied to the health education literature when reporting and interpreting effect sizes. Guidelines include: (1) expressing what effect sizes are being reported; (2) interpreting the effect sizes by taking into consideration both their assumptions and their limitations; and (3) reporting confidence intervals for effect sizes and other study results. First, researchers must *say exactly what effect sizes are being reported.* Due to the numerous effect size choices, the reader cannot accurately evaluate the effect if he or she does not know which effect size to interpret. Specifying the reported effect size also allows readers to convert the estimates so that they may be expressed in the same metrics. For example, if a reader is reviewing two articles, one where the author reported Cohen's d and the other where the author reported Pearson's r, he or she can convert both estimates to reflect either Cohen's d or Pearson's r. In this manner, "apples-to-apples"[14] comparisons can be made.

Second, researchers should *interpret effect sizes by taking into consideration both their assumptions and their limitations.* When analytical assumptions are violated, the results and effect estimates are compromised.[14] For example, when comparing effect size results across studies, researchers must consider differences in study designs. Effect sizes have a connection to the designs that they support, so acknowledging their differences is vital to understanding their estimates in relation to other studies. Among readers who desire to perform meta-analyses, reporting effect sizes will be especially useful. However, it is mindful to consider that future researchers may choose to disregard a study that does not include effect sizes simply because they cannot find a comparable way to estimate the desired effect size results.

Third, researchers should *report confidence intervals for effect sizes and other study results.* Confidence intervals are easily modifiable to a graphical representation of the data, allowing a number of studies to be depicted efficiently.[22] As a highly recommended technique, the widths of confidence intervals can be compared to evaluate the precision of the estimates in a given study.[22] Confidence intervals used to estimate a population value generally are symmetrical or nearly symmetric around a value. Due to their tendency to convey information about the precision of an estimated population value as well as statistical significance, confidence intervals are preferable to p values. Readers are strongly encouraged to view other resources[14,22] for strategies used to construct confidence intervals.

CONCLUSION

This article demonstrates that reporting and interpreting effect sizes in health education journals serves multiple purposes. Effect sizes provide easier interpretation of results for health educators as well as provide future researchers with an understanding of the strengths of the associations between the variables mentioned.[22] Additionally, effect sizes increase the understanding and readability of results, determine practical significance, and enhance the discipline by providing information for statistical comparison across studies. Despite these benefits, health education journals currently provide minimal instructions regarding statistical analysis reporting. Specifically, there are no requirements to include effect sizes, confidence intervals for effect sizes, or graphics in the author submission guidelines. If health education desires to enhance its position as a science, it must embrace statistical reform. Implementing effect size reporting and interpretation will contribute to the success of this pursuit.

REFERENCES

1. American Psychological Association. *Publication Manual of the American Psychological Association.* 5th ed. Washington, D.C.: Author; 2001.

2. Merrill RM, Stoddard J, Shields EC. The use of statistics in the American Journal of Health Education from 1994 through 2003: A content analysis. *American Journal of Health Education.* 2004; 35:290–297.

3. Buhi E. The Insignificance of "Significance" Tests: Three recommendations for health education researchers, *American Journal of Health Education.* 2005; 36: 109–112.

4. Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Stat Methods in Med Research.* 2000; 9:41–55.

5. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Researcher.* 1976;5(10):3–8.

6. Thompson B, ed. *Score reliability: Contemporary thinking on reliability issues.* Newbury Park, CA: Sage; 2003.

7. Kirk RE. Practical significance: A concept whose time has come. *Educ and Psychological Meas.* 1996;61:213–218.

8. Trusty J, Thompson B, Petrocelli JV. Practical Guide for Reporting Effect Sizes in Quantitative Research in the Journal of Counseling & Development. *J of Counseling and Dev.* 2004;82:107–110.

9. Elmore F. *A primer on basic effect size concepts.* Paper presented at the annual meeting of the American Educational Research Association. 2001, April. Seattle, WA.

10. Huberty CJ. A history of effect size indices. *Educ and Psych Meas.* 2002;62(2):227–240.

11. Capraro RM, Capraro M. Treatments of effect sizes and statistical significance tests in textbooks. *Educ and Psych Meas.* 2002;62:771–782.

12. Wilkinson L. Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *Am Psychologist.* 1999;54:594–604.

13. Thompson B. Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educ Psych Review.* 1999; 11:157–169.

14. Vacha-Haase T, Thompson B. How to estimate and interpret various effect sizes. *J of Counseling Psych.* 2004; 51(4):473–481.

15. Thompson B. Oneway Analysis of Variance (ANOVA). In: *Foundations of Univariate Behavioral Statistics.* New York, New York: Guilford Press; In press: 1–25.

16. Thompson B. *Canonical correlation analysis.* In L. Grimm & P. Yarnold (Eds.), Reading and understanding more multivariate statistics. Washington, D.C.: American Psychological Association. 2000:285–316.

17. Thompson B. Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply Here: A Guidelines Editorial, *Educational and Psych Measurement.* 1995;55(4):525–534.

18. Kline R. Beyond statistical significance testing: reforming data analysis methods in behavioral research. Washington, D.C.: American Psychological Association; 2004.

19. Cohen J. Multiple regression as a general data-analytic system. *Psych Bulletin.* 1968;70:426–443.

20. Thompson B. Significance, effect sizes, stepwise methods, and other issues: Strong arguments more the field. *J of Experimental Educ.* 2001;70:80–93.

21. Gage NL. Instructions for authors Journal of Educational Psychology. *J of Educ Psych.* 2003;95:201.

22. Thompson B. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educ Researcher.* 2002;31(3):24–31.