The Incorporation of Effect Size in Information Technology, Learning, and Performance Research

Joe W. Kotrlik

Heather A. Williams

Research manuscripts published in the Information Technology, Learning, and Performance Journal *are expected to adhere to the publication guidelines of the American Psychological Association (2001) and generally accepted research and statistical methodology. This manuscript describes the rationale supporting the reporting of effect size in quantitative research and also provides examples of how to calculate effect size for some of the most common statistical analyses. We include a table of recommendations for effect size interpretation. We also address basic assumptions and cautions on the reporting of effect size.*

Introduction

Effect size is a term used to describe a family of indices that measure the magnitude of a treatment effect. Effect size is different from significance tests because these measures focus on the meaningfulness of the results and allow comparison between studies, furthering the ability of researchers to judge the practical significance of results presented. It is almost always necessary to include some index of effect size or strength of relationship in your results section so that the consumers of your research will be able to truly understand the importance of your findings and make comparisons among studies.

A review of manuscripts in the *Information Technology, Learning, and Performance Journal* from the past three years revealed that effect size has not been consistently reported. Out of the 16 manuscripts published from 1999-2001, effect size should have been reported in 11 manuscripts. An analysis of these 11 manuscripts revealed that effect size was reported correctly in one manuscript, not reported at all in six manuscripts, and reported incorrectly or inappropriately in the remaining four manuscripts.

The need to improve the reporting of effect size is an issue being addressed by numerous journals. This article will address the use and reporting of effect size in *Information Technology, Learning, and Performance Journal* manuscripts.

Research/Theoretical Base

What is Effect Size?

The term effect size has become increasingly popular throughout educational literature in recent decades; and even more prevalent in recent years. The APA Task Force emphasized that researchers should "always provide some effect-size estimate when reporting a *p* value. . . . reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). Kirk (1996) also emphasized the importance of the effect size concept.

Joe W. Kotrlik is Professor, School of Human Resource Management and Workforce Development, Louisiana State University, Baton Rouge, Louisiana.

Heather A. Williams is performance improvement consultant, Reserve, Louisiana and a recent graduate, School of Human Resource Education and Workforce Development, Louisiana State University, Baton Rouge, Louisiana.

With these discussions focusing on effect size, the literature presents a variety of effect size definitions. Common definitions include: a standardized value that estimates the magnitude of the differences between groups (Thomas, Salazar, & Landers, 1991), a standardized mean difference (Olejnik & Algina, 2000; Vacha-Haase, 2001), the degree to which sample results diverge from the null hypothesis (Cohen, 1988, 1994), a measure of the degree of difference or association deemed large enough to be of 'practical significance' (Morse, 1998), an estimate of the degree to which the phenomenon being studied exists in the population (e.g., a correlation or difference in means) (Hair, Anderson, Tatham, & Black, 1995), and strength of relationship (American Psychological Association, 2001).

Among the numerous effect size definitions, the majority of definitions include the descriptors "standardized difference between means" or "standardized measure of association." In practice, researchers commonly use two categories of measures of effect size in the literature, namely, measures of effect size (according to group mean differences), and measures of strength of association (according to variance accounted for) (Maxwell & Delaney, 1990).

Why is Effect Size Important?

An early discussion of effect size by Karl Pearson (1901) addressed the idea that statistical significance provides the reader with only part of the story and therefore must be supplemented. Fisher (1925) followed up on this discussion by proposing that researchers present correlation ratios or measures of the strength of association when reporting research findings. Since that time, numerous researchers have argued for the use of effect size statistics to complement or even replace statistical significance testing results, allowing the reader to interpret the results presented as well as providing a method of comparison of results between studies (Cohen, 1965, 1990, 1994; Hays, 1963; Kirk, 1996, 2001; Thompson, 1998, 2002). Effect size can also characterize the degree to which sample results diverge from the null hypothesis (Cohen, 1988, 1994). Therefore, reporting effect size allows a researcher to judge the magnitude of the differences present between groups, increasing the capability of the researcher to compare current research results to previous research and to judge the practical significance of the results derived.

Today, the most widely referenced impetus for authors to report effect size is the following statement printed in the 2001 *Publication Manual of the American Psychological Association*:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. . . . 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. (p. 25-26)

These statements followed the American Psychological Association Task Force's earlier recommendation strongly urging authors to report effect sizes such as Cohen's *d*, Glass's *delta*, eta^2, or adjusted R^2 (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599).

Support for this direction by such a prominent group is witnessed throughout the literature. For example, Fan (2001) described good research as presenting both statistical significance testing results as well as effect sizes. Baugh (2002) declared, "Effect size reporting is increasingly recognized as a necessary and responsible practice" (p. 255). It is a researcher's responsibility to adhere to the most stringent analytical and reporting methods possible in order to ensure the proper interpretation and application of research results.

Effect Size versus Statistical Significance

So why the confusion and what is the difference between statistical significance testing and effect sizes? The confusion stems from misconceptions of what statistical significance testing tells us. First, as Nickerson (2000) explains, there is a belief that a small value of *p* means a treatment effect of large

magnitude, and that statistical significance means theoretical or practical significance. "Statistical significance testing does not imply meaningfulness" (Olejnik & Algina, 2000, p. 241). Statistical significance testing evaluates the probability of obtaining the sampling outcome by chance, while effect size provides some indication of practical meaningfulness (Fan, 2001). Statistical significance relies heavily on sample size, while effect size assists in the interpretation of results and makes trivial effects harder to ignore, further assisting researchers to decide whether results are practically significant (Kirk, 2001).

Table 1: Percentage of Time College Faculty Actually Spent Teaching Compared to Percentage of Time College Faculty Preferred to Spend Teaching

Time spent	Actual %		Preferred %		Comparison			
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	Cohen's *d*
Teaching	52.71	33.86	49.74	28.57	2.20	154	.03	.09

Note. Information in table taken from Williams, H. A. (2003).

Arguments surrounding the frequent misuse of statistical significance testing populate education and related literature. One argument is that statistical significance is a function of sample size and, therefore, the larger the sample size, the more likely it is that significant results will occur (Fan, 2001; Kirk, 1996; Thompson, 1996). Fan (2001) furthers this argument by positing that the large reliance on statistical significance testing "often limits understanding and applicability of research findings in education practice" (p. 275). A second argument is that researchers often associate statistical significance with practical significance (Daniel, 1998; Nickerson, 2000). A third argument states that it is the researcher who inappropriately uses and interprets this statistical methodology and, therefore, it is the researcher and not the method that is the source of the deficiency (Cortina & Dunlap, 1997; Kirk, 1996).

An example in a recent study demonstrates why the reporting of effect size is important. In this example, failing to report effect size would have resulted in the researcher using the results of a statistically significant t-test to conclude that important differences existed. In this study, Williams (2003) compared the percent of time that faculty actually spent in teaching with the percent of time they preferred to spend in teaching. The data in Table 1 show that even though the t-test was statistically significant ($t = 2.20$, $p = .03$, $df = 154$), Cohen's effect size value ($d = .09$) did not meet the standard of even a "small" effect size. This indicated that the difference has low practical significance, so Williams made no substantive recommendations based on the results of this *t*-test.

When one considers the discussion surrounding effect size in the literature, it is apparent that it is the researcher's responsibility to understand the test results and properly use the reporting methodology which best represents the true findings of the research effort. This responsibility increases in importance because, although a statistically significant outcome may or may not be practically meaningful, a practically meaningful outcome may or may not have occurred by chance (Fan, 2001). These possibilities add challenge to the presentation of results, leading to the recommendation that a researcher should present both the statistical significance test results and an appropriate effect size measure (Carver, 1993; Fagley & McKinney, 1983; Robinson & Levin, 1997; Shaver, 1993; Thompson, 1996, 1998).

Basic Assumptions and Cautions on the Use of Effect Size Interpretation

We have established the justification for the use of effect sizes. There are a few issues that a researcher must keep in mind, however, when choosing an effect size measure and reporting that measure, aside from which measure goes best with the type of significance test used. First, a researcher must determine the appropriate statistical test to analyze the data. This should be based on sampling assumptions, sample design, and research objectives. Following the determination of the most appropriate significance test, the researcher should

decide how the effect size should be calculated and should report information on both the selection of the analytical method and the calculation and selection of the effect size measure. The consistent application of these two steps will further improve the quality of research reporting in the *Information Technology, Learning, and Performance Journal*.

Another issue concerns how effect size is described or interpreted. Selecting the set of descriptors to use for effect size interpretations for correlation coefficients, for example, is similar to deciding the statistical significance level. Just as a researcher might have a set of assumptions or other logical reasons that were used to choose an *alpha* level of .01 rather than .05, a researcher may decide to use a more conservative set of descriptors when he or she desires more confidence in the declaration of the existence of practical relationships. The set of descriptors by Davis (1971), for example, describes a correlation of .10 as having a low relationship, while Hinkle, Wiersma, and Jurs (1979) describes this same relationship as "negligible." Authors should cite the basis for their method of reporting effect sizes.

As a final caution, this manuscript provides a rather basic discussion of the appropriate use of effect sizes. It is the researcher's responsibility to investigate the appropriate use of effect sizes in his or her individual research efforts.

Measures of Effect Size

Although Cohen (1988) and other authors did not anticipate that their scales for the interpretation of effect size magnitude would be so widely used and accepted, these scales have become a required facet of quality research. The acceptance of the use of effect size interpretation scales permeates education literature. A word of caution–it is the researcher who is the closest to the data and the design of the study; therefore, it is the researcher who should describe the magnitude of the results based on the study itself and previous research in that area. It is the researcher's responsibility to choose the appropriate method with which to describe the magnitude of an effect size. We present below methods for determining effect size for the most commonly used statistical analyses.

Parametric Correlations (Simple and Multiple)

Perhaps the simplest measures and reporting of effect sizes exist for correlations. The correlation coefficient itself is a measure of effect size. The most commonly used statistics for parametric correlations are Pearson's *r*, Spearman's *rho*, and point bi-serial. The size of these correlation coefficients must be interpreted using a set of descriptors for correlation coefficients as described in Table 2.

Regression (Simple and Multiple)

Another simple measure of effect size is the multiple regression coefficient, R^2. All standard statistical analysis packages calculate this coefficient automatically, and this coefficient represents the proportion of variance in the dependent variable explained by the independent variable(s). The effect size for this coefficient may be interpreted using a set of descriptors derived from Cohen's f^2 statistic (see Table 2).

Independent Samples t-tests

One method to estimate effect size for independent samples t-tests is to calculate and interpret Cohen's *d* statistic (Cohen, 1988). If your statistical analysis program does not calculate Cohen's *d*, you will need three pieces of information to calculate this statistic: the means of the two samples and the pooled variance estimate. If your program does not calculate Cohen's *d*, it probably does not provide the pooled standard deviation either. Use the following formulas to calculate Cohen's *d*:

$$\text{Pooled standard deviation} = \sqrt{\left[(n_1-1)s_1^2+(n_2-1)s_2^2\right]/\left[(n_1-1)+(n_2-1)\right]}$$

$$\text{Then, Cohen's } d = \frac{\text{Difference between sample means}}{\text{Pooled standard deviation}}$$

After calculating Cohen's *d*, use the descriptors in Table 2 to interpret this coefficient.

Table 2: Descriptors for Interpreting Effect Size

Source	Statistic	Value	Interpretation
Rea & Parker, 1992	Phi or Cramér's V	.00 and under .01	Negligible association
		.10 and under .20	Weak association
		.20 and under .40	Moderate association
		.40 and under .60	Relatively strong association
		.60 and under .80	Strong association
		.80 and under 1.00	Very strong association
Cohen, 1988	Cohen's d	.20	Small effect size
		.50	Medium effect size
		.80	Large effect size
	Cohen's f	.10	Small effect size
		.25	Medium effect size
		.40	Large effect size
Cohen, 1988	R^2	.0196	Small effect size
		.1300	Medium effect size
		.2600	Large effect size
Kirk, 1996	Omega squared	.010	Small effect size
		.059	Medium effect size
		.138	Large effect size
Davis, 1971[a]	Correlation coefficients	.70 or higher	Very strong association
		.50 to .69	Substantial association
		.30 to .49	Moderate association
		.10 to .29	Low association
		.01 to .09	Negligible association
Hinkle, Wiersma, & Jurs, 1979[a]	Correlation coefficients	.90 to 1.00	Very high correlation
		.70 to .90	High correlation
		.50 to .70	Moderate correlation
		.30 to .50	Low correlation
		.00 to .30	Little if any correlation
Hopkins (1997)[a]	Correlation coefficients	.90 to 1.00	Nearly, practically, or almost: perfect, distinct, infinite
		.70 to .90	Very large, very high, huge
		.50 to .70	Large, high, major
		.30 to .50	Moderate, medium
		.10 to .30	Small, low, minor
		.00 to .10	Trivial, very small, insubstantial, tiny, practically zero

[a] Several authors have published guidelines for interpreting the magnitude of correlation coefficients. Three of those sets of guidelines are provided in this table.

Paired Samples t-tests

To calculate effect size for paired t-tests, Cohen's *d* statistic may be calculated. If your statistical analysis program does not calculate Cohen's *d*, you will need two pieces of information to calculate this statistic: the mean of the differences between the pairs and the standard deviation of the differences. Use the following formula to calculate Cohen's *d*:

$$\text{Cohen's } d = \frac{\text{Mean difference between the pairs}}{\text{Standard deviation of the differences}}$$

After calculating Cohen's *d*, use the descriptors in Table 2 to interpret this coefficient.

Analysis of Variance

Two methods of reporting effect sizes for analyses of variance are Cohen's *f* (Cohen, 1988) and omega squared (Ω^2). Both provide an estimate of the proportion of variance explained by the categorical variable, with Cohen's *f* estimating the proportion of variance explained for the sample, while Omega squared estimates the proportion of variance explained for the population.

To calculate Cohen's *f*, eta squared (ω^2) must be calculated as follows:

$$w^2 = SS_{Between}/SS_{Total}$$

Then, use the following formula to calculate Cohen's *f*:

$$\text{Square root of } (w^2/1-w^2)$$

Calculate omega squared (Ω^2) as follows:

$$W^2=\frac{SS_{Between}-(k-1)MS_{Within}}{SS_{Total}+MS_{Within}} \quad (k=\text{number of groups})$$

The sum of square and mean square information is provided by most statistical programs. After calculating Cohen's *f* or omega squared, use the descriptors in Table 2 to interpret these coefficients.

Contingency Tables

In estimating the magnitude of association in contingency tables, use the phi coefficient for a 2 x 2 table. Phi is a Pearson product-moment coefficient calculated on two nominal, dichotomous variables when the categories of both variables are coded 0 and 1. To describe the magnitude of association between categorical variables for a contingency table larger than 2 x 2, use Cramér's V. Many statistical analysis programs, including SPSS and SAS, will automatically calculate either the phi or Cramér's V coefficients. Use the descriptors in Table 2 to interpret these coefficients.

Conclusion

The effect size measures described are not intended to represent all effect size measures available. Effect size measures were presented only for the most commonly used statistical analyses. The references provided at the end of this article will serve as a good starting point for authors attempting to identify appropriate measures of effect size for other types of statistical analyses.

The authors' main purpose for writing this article was their hope that *Information Technology, Learning, and Performance Journal* authors will improve the reporting of their research by including effect size measures when appropriate. We hope that this article will serve as a resource for researchers.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62(2),* 254-263.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology.* New York: Academic Press.

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

Cohen. J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

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

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

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and mis-interpretation with implication for the editorial policies of educational journals. *Research in the Schools, 5(2),* 23-32.

Davis, J. A. (1971). *Elementary survey analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Fagley, N. S., & McKinney, I. J. (1983). Review bias for statistically significant results: A reexamination. *Journal of Counseling Psychology, 30,* 298-300.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research (Washington, D.C.), 94(5),* 275-282.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Hair, J. F, Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (4th ed.). NJ: Prentice Hall.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart & Winston.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (1979). *Applied statistics for the behavioral sciences*. Chicago: Rand McNally College Publishing.

Hopkins, W. G. (1997). New view of statistics. Retrieved August 23, 2002 from http://www.sportsci.org/resource/stats/effectmag.html

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

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61(2),* 213-218.

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

Morse, D. T. (1998). MINSIZE: A computer program for obtaining minimum sample size as an indicator of effect size. *Educational and Psychological Measurement, 58(1),* 142-154.

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

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241-286.

Pearson, K. (1901). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of The Royal Society of London, 195,* 1-47.

Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research*. San Francisco: Jossey-Bass.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26(5),* 21-26.

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

Thomas, J. R., Salazar, W., & Landers, D. M. (1991). What is missing in $p < .05$? Effect size. *Research Quarterly for Exercise and Sport, 62(3),* 344-348.

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

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

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

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61(2),* 219-224.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Williams, H. A. (2003). A mediated hierarchical regression analysis of factors related to research productivity of human resource development postsecondary faculty (Doctoral Dissertation, Louisiana State University, 2003). Available: http://etd01.lnx390.lsu.edu:8085/docs/available/etd-0326103-212409/ (Retrieved April 9, 2003).