

Do Effect-Size Measures Measure Up?: A Brief Assessment

Anthony J. Onwuegbuzie¹

University of South Florida

Joel R. Levin

University of Arizona

Nancy L. Leech

University of Colorado at Denver

Because of criticisms leveled at statistical hypothesis testing, some researchers have argued that measures of effect size should replace the significance-testing practice. We contend that although effect-size measures have logical appeal, they are also associated with a number of limitations that may result in problematic interpretations of them in research on children and adults with learning disabilities (LD). The purpose of the present paper is to provide a framework for reporting and interpreting empirical research findings in LD research. Specifically, we recommend that: (1) researchers apply criteria of both statistical significance and substantive significance to help consumers of research assess the believability and importance of reported results, respectively; with (2) the establishment of statistical significance, obtained via the use of inferential statistical techniques serving as a precursor to the interpretation of measures of substantive importance. We further contend that the family of standard effect-size indices represents just one approach for assessing substantive significance in LD research. Other methods include the use of confidence intervals and consideration of the results' clinical significance and economic significance. In addition, the critical role played by independent replications must not be overlooked by LD researchers. As such, effect-size measures have an important, though not exclusive, function in evaluating educational and psychological research findings in general and LD research results in particular.

One of the greatest challenges in the area of learning disabilities (LD) research is that researchers present sufficient information regarding their findings to enable consumers of LD research to make effective, data-driven decisions. Unfortunately, at present, too many researchers in the field of LD and elsewhere provide insufficient information, which prevents consistency between a researcher's data analysis and interpretations from occurring (Levin & Robinson, 2000). In particular, these researchers tend to report only one kind of significance index (e.g., p-value, effect-size index), possibly resulting in misleading conclusions being drawn. Thus, the purpose of the present paper is to provide a framework for reporting and interpreting empirical findings in LD research. Although our recommendations in this paper are targeted at LD researchers, they are applicable to research conducted in all other areas of education and psychology.

SETTING THE SCENE

The last eight decades have witnessed fervent attacks launched on statistical hypothesis testing in all areas of research, including the area of learning disabilities (LD). For example, Thompson (1993) outlined the following three major limitations of statistical hypothesis testing: (a) overreliance on sample size; (b) certain meaningless comparisons; and (c) unavoidable dilemmas created by statistical significance testing, such as testing for model assumption versus testing the research hypothesis. Referring to the practice of statistical hypothesis testing, Meehl (1978, p. 817) went so far as to state that it "is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." Thus, researchers, including LD researchers, have been under attack for their use of statistical hypothesis testing, many without knowing it was occurring.

Because of the limitations of statistical hypothesis testing posited by its opponents, an increasing number of researchers have been advocating the calculation, reporting, and interpretation of measures of effect size either to supplement or to replace statistical significance testing. In particular, the most recent edition of the pivotal *Publication Manual of the American Psychological Association* (2001) states:

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. (pp. 25-26)

Further, at the time of writing this paper, the editorial policies of 23 journals, which in total attract tens of thousands of readers, currently require effect-size reporting. It is expected that other journals will "jump on the bandwagon" in the upcoming months.

DEFINITION OF AN EFFECT SIZE

An effect size represents a label given to a family of indices that measure the magnitude of a difference or a relationship. More specifically, Cohen (1988) defines an effect size as follows:

Without intending any necessary implication of causality, it is convenient to use the phrase "effect size" to mean "the degree to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false." By the above route it can now readily be clear that when the null hypothesis is false, it is false to some specific degree, i.e., the effect size (ES) is some specific non-zero value in the population. The larger this value, the greater the degree to which the phenomenon under study is manifested. (pp. 9-10)

Kirk (1996) has identified 61 different effect-size indices, while

¹ Correspondence should be addressed to Anthony J. Onwuegbuzie, Department of Educational Measurement and Research, College of Education, University of South Florida, 4202 East Fowler Ave, EDU 162, Tampa, Florida 33620-7750, or E-Mail: tonyonwuegbuzie@aol.com

Huberty and his colleagues (e.g., Huberty & Lowman, 2000) have developed new indices of effect size that they call Group Overlap indices. All of these effect sizes can be classified into two broad categories: (a) measures of standardized differences, also known as “*d* family” effect-size indices (e.g., Cohen’s *d*, Glass’s Δ , Hedges’ *g*) and (b) variance-accounted-for measures, also known as “*r*² family” effect-size indices (e.g., *r*², *R*², η^2 , ω^2)—see, for example, Majova-Seane (2003). Sample effect-size measures also can be classified as being “uncorrected” or “corrected” indices, or as representing more or less biased estimators of their population counterparts (Kirk, 1996; Olejnik & Algina, 2000). For example, in multiple regression, researchers can compute *R*² (uncorrected effect size) and/or adjusted *R*² (corrected effect size). Similarly, for the ANOVA family, the sample η^2 and ω^2 represent, respectively, more and less biased estimators (Cohen, 1988).

An example of one standardized difference index. One of the most common standardized effect-size measures, Cohen’s *d*, is used to indicate the magnitude of the difference between two groups (Cohen, 1988). In the two independent-samples case, each group’s mean, standard deviation, and sample size can be used to calculate *d* according to the following formula.

$$d = \frac{M_1 - M_2}{SD_{pooled}} \quad (1)$$

where *M*₁ is the mean of one group, *M*₂ is the mean of the other group, and *SD*_{pooled} is the pooled standard deviation (the square root of the pooled variance). Alternatively, *d* in the same context can be calculated from the obtained *t* statistic for assessing the group mean difference, as:

$$d = t \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \quad (2)$$

where *N*₁ and *N*₂ represent the respective group samples sizes. When *N*₁ = *N*₂ = *n*, this simplifies to:

$$d = t \sqrt{\frac{2}{N_1}} \quad (3)$$

The *d* is expressed in standard deviation units, which helps with the interpretation. Therefore, for instance, a *d* of 1.00 indicates that the group means differ by one pooled within-group standard deviation. Cohen’s *d* is particularly useful in research **involving learning-disabled (LD) participants, wherein** two groups often are compared. For example, *d* could be used to assess the standardized difference in levels of academic motivation between children with LD and children without LD.

An example of proportion of variance explained. Many effect-size measures are expressed as the proportion of dependent-measure variance accounted for, or “explained,” by the independent variable. One of the most common is *r*². The *r*² gives information about the strength of linear association or relationship between variables of interest. Thus, for example, an *r*² of .70 indicates that 70% of the variance in the dependent variable (*Y*) can be explained by the independent variable’s (*X*’s) linear relationship with *Y*.

ADVANTAGES OF EFFECT-SIZE ESTIMATES

Proponents of effect sizes contend that they provide many advantages. Moreover, Wilkinson and the Task Force on

Statistical Inference (1999, p. 599) claim that “reporting and interpreting effect sizes in the context of previously reported effects is essential to good research.” Further, Thompson (2002) declared that the potential benefits of reporting and interpreting an effect size “arise not from interpreting effects against benchmarks, but rather by comparing effect sizes directly with the effects reported in related prior studies” (p. 30).

Another attribute of effect-size measures highlighted by proponents is that these indices can be converted into each other’s metrics. Specifically, measures of standardized differences (i.e., *d*) can be converted to variance-accounted-for measures (i.e., *r*²)—cf. Cohen (1988). Further, it is argued that placing confidence intervals around effect size estimates promotes meta-analytic thinking, which involves “the thoughtful integration of all prior related research when formulating expectations and designing a study” (Thompson, 2002, p. 28).

LIMITATIONS OF EFFECT SIZES

Because effect-size measures offer the potential for increasing our understanding of phenomena in LD research and beyond, some researchers (e.g., Carver, 1993) have recommended that statistical hypothesis testing be banned completely and be replaced by effect-size estimates. Unfortunately, we contend that such thinking is not justified because effect sizes are subject to as much abuse and misuse as are tests for statistical significance. Disturbingly, however, whereas the limitations of statistical hypothesis testing have been identified and published extensively since 1919 (e.g., Boring, 1919; Carver, 1993; Kirk, 1996; Thompson, 1993), problems associated with the use of effect-size measures have rarely been communicated. That is, although effect sizes have logical appeal, it is clear that their use has not been subjected to the same degree of scrutiny as has statistical hypothesis testing. As such, it appears that many researchers are unaware that certain of the same criticisms leveled at statistical hypothesis testing also can be directed at effect-size estimation.

Recently, however, Onwuegbuzie and Levin (in press) provided an extensive critique of effect-size indices. In particular, these authors identified nine major concerns and limitations of effect-size estimation. In particular, effect-size indices are subject to the following limitations: (a) effect sizes can vary as a function of one’s research objective (i.e., theory application or effects application); (b) effect sizes can vary as a function of one’s research design and experimental conditions; (c) researchers can select from a variety of effect-size measures to argue different (possibly self-serving) points; (d) guidelines for interpreting effect size magnitudes are inconsistent and generally arbitrary; (e) effect sizes can vary as a function of sample size and sample variability; (f) effect sizes are sensitive to departures from normality; (g) effect sizes can vary as a function of the variability of the outcome measure (both between and within samples); (h) effect sizes can vary as a function of the reliability of the outcome measure; and (i) effect sizes can vary as a function of the scale of measurement used (i.e., nominal, ordinal, interval, ratio).

However, perhaps an even more troubling problem associated with measures of effect sizes that was not mentioned by Onwuegbuzie and Levin (in press) is that such measures are not always meaningful or useful for consumers of research. As

was pointed out by Leech and Onwuegbuzie (2003), “[I]t is difficult for a policymaker to know how to act upon a Cohen’s *d* effect size of .28, even if confidence limits are placed around this value.” Moreover, as was noted by these authors, there are times when other indicators of substantive significance are more appropriate. One such alternative, clinical significance, refers to the extent that the intervention makes a real difference to the quality of life of the study participants or to those with whom they come into contact (Kazdin, 1999). This “clinical significance” criterion is reflected in an example provided by Levin (1998, pp. 45-46) and represents a related distinction made by Levin (2002) – one that encourages research consumers to judge the importance of a study’s outcomes in two complementary ways:

(a) “internally,” or in relation to the specific study’s characteristics (including the participant sample), for which a standardized-effect measure...is evaluated; and (b) “externally,” in relation to extra-study criteria, where other normative measures of educational consequence or impact may be taken into account. (p. 486)

A second substantive-significance alternative, economic significance, refers to the economic value of the effect of a treatment or intervention, and consists of the following five indicators: cost effectiveness, cost benefit, cost utility, cost feasibility, and cost sensitivity (Leech & Onwuegbuzie, 2003). For example, rather than reporting that: “The reading intervention improved LD students’ average reading scores by $d = .83$,” it would likely be much more useful for consumers to learn that: “The students’ average reading scores improved from 2.4 years below grade level to only 0.3 years below grade level [clinical significance relative to an external criterion]. Moreover, factoring in a yearly intervention expenditure of \$122 per student, this represents an average cost savings (in terms of school special education services and out-of-school remedial instruction) of \$236 per student per year [economic significance].” These additional indicators imply that the family of effect-size measures represents just one approach for assessing substantive significance and that researchers need to pay attention to the clinical and societal implications of their work as well.

A *sequential rapprochement*. That both statistical significance testing and effect size estimation have limitations provide credence to earlier (e.g., Levin, 1993; Robinson & Levin, 1997) assertions that these two sets of procedures should be used in combination. Those authors have proposed a sequential (“two-step”) strategy for analyzing primary-research data, in which an effect size is reported and interpreted if and only if a statistical test applied to the data indicates that the associated outcome of interest is not likely due to chance. As such, hypothesis testing serves as a critical academic “gatekeeper, guarding against spurious effect-size estimation” (Robinson & Levin, 1997, p. 13). Encouragingly, the two-step approach helps reduce not only the problem of effect-size misrepresentation but also “significance” testing’s problem of *p*-value misinterpretation. In this respect, the sequential combination of statistical hypothesis testing and effect-size estimation serves to increase what Levin and Robinson (2000) call “conclusion coherence,” or consistency between a researcher’s data analysis and conclusions.

At the same time, we heartily endorse the use of conven-

tional confidence intervals **in LD research** because they “kill two birds (hypothesis testing and estimation) with one stone (a single inferential statistical procedure).” Specifically, constructing confidence intervals based on an *a priori* significance level (Levin, 1998; Onwuegbuzie & Levin, in press) allows researchers to conduct statistical tests of the researcher’s hypothesis (Krantz, 1999) while at the same time providing a range of magnitudes for the effect size (Abelson, 1997). In addition, the second part (effect-size estimation) of Robinson and Levin’s (1997) two-step approach can easily be adapted to incorporate the above-mentioned notions of clinical and economic significance in LD research.

SUMMARY AND CONCLUSIONS

Because of criticisms launched at statistical significance testing, many researchers have called for the reporting of measures of effect size, either in addition to or instead of testing for statistical significance. Unfortunately, proponents of effect-size measures tend to focus on their strengths and not on their limitations, thereby suggesting that effect sizes represent a panacea for educational and psychological research, including research conducted in LD populations. Based on the effect-size limitations mentioned here and elsewhere, however, the panacea may be more imagined than real.

Nevertheless, we believe that measures of effect size should play a prominent role in LD research – as long as LD researchers incorporate them coherently and are cognizant of their limitations. Further, when reporting and interpreting effect sizes, LD researchers should specify the underlying assumptions and delineate as many design, psychometric, and analysis characteristics as possible so that these estimates can be contextualized (Onwuegbuzie & Levin, in press). We hope that the concept of substantive significance is broadened beyond “internally referenced” effect sizes to encompass “externally referenced” indicators representing clinical significance, economic significance, and the like. Finally, and most importantly, we strongly encourage LD researchers to conduct both literal and constructed independent replication studies (Lykken, 1968)—either within the context of an individual research report or across multiple reports—as undoubtedly the single best “procedure” for yielding dependable research findings and new knowledge

Anthony J. Onwuegbuzie, Ph.D., is an associate professor in the Department of Educational Measurement and Research at the University of South Florida in Tampa, Florida, where his research primarily involves disadvantaged and under-served populations such as minorities, learning disabled students, and juvenile delinquents. Joel R. Levin, Ph.D., is a professor in the Department of Educational Psychology at the University of Arizona. His areas of expertise include the design and analysis of educational research as well as cognitive strategies and processes. Nancy L. Leech, Ph.D., is an assistant professor in the Educational Psychology Division of the School of Education at the University of Colorado at Denver. Her research topics include willingness to seek counseling, mentoring in higher education, and gender and equity issues.

REFERENCES

- Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 117-141.
- American Psychological Association. (2001). *Publication manual of the*

- American Psychological Association* (5th ed.). Washington, DC: Author.
- Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin*, 16, 335-338.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287-292.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Wiley.
- Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, 60, 543-563.
- Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 332-339.
- Kirk, R. E. (1996). Practical significance. A concept whose time has come. *Education and Psychological Measurement*, 56, 746-759.
- Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 94, 1372-1381.
- Leech, N. L., & Onwuegbuzie, A. J. (2003, April). *A proposed fourth measure of significance: The role of economic significance in educational research*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, 61, 378-382.
- Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools*, 5(2), 43-53.
- Levin, J. R. (2002). How to evaluate the evidence of evidence-based interventions? *School Psychology Quarterly*, 17, 483-492.
- Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, 29(1), 34-36.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159.
- Majova-Seane, N. (2003, February). *The inclusion of effect sizes in addition to statistical significance testing reporting*. Paper presented at the annual meeting of the Southwestern Educational Research Association, San Antonio, TX.
- 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.
- Olejnik, S., & Algina, J. (2000). Measures of effects size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241-286.
- Onwuegbuzie, A. J., & Levin, J. R. (in press). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods*.
- Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 26(5), 21-26.
- Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. *The Journal of Experimental Education*, 61, 361-377.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 25-32.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.



Sharpen your skills for students with dyslexia, ADHD and other learning difficulties.

A higher caliber of educational professionals than ever before is needed to meet the growing number of individuals struggling with learning disabilities. **The National Institute for Learning Disabilities (NILD)** trains qualified applicants as educational therapists to meet such needs through a series of three graduate level courses.

NILD Educational Therapy™ is a successful model of intervention for children and adults with learning disabilities. Many students are now achieving academic levels commensurate with their IQ and are successful independent learners. Across the nation, educators are understanding the reasons for students' perplexing difficulties in learning.

Discover NILD's integrative program of intervention, providing students a chance to succeed. Join us in March 2004, for our annual educational conference in the Boston area. This conference will provide an overview of the NILD model and will target specific topics relating to a better understanding of learning disabilities.

Mention this ad and receive a \$25 conference discount.

NILD Course Descriptions:

Level I expands understanding of the field of learning disabilities and provides foundational training in the philosophy and techniques of NILD Educational Therapy™.

Level II provides a review of basic NILD Educational Therapy™ techniques and introduces further techniques for students.

Level III provides in-depth study of the neurological considerations of NILD Educational Therapy™ techniques and further develops clinical and professional skills.

Call 800-616-6453, email info@mail.nild.net or visit our website www.nild.net today for information on becoming an educational therapist in a program that has successfully served students for over 20 years.



Copyright of Learning Disabilities -- A Contemporary Journal is the property of Learning Disabilities Association of Massachusetts and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Copyright of Learning Disabilities -- A Contemporary Journal is the property of Learning Disabilities Association of Massachusetts and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.