Chapman F6

Viewpoint

Effect-Size Reporting Practices in *AJSLP* and Other ASHA Journals, 1999–2003

Timothy Meline
Bailey Wang
The University of Texas–Pan American, Edinburg

A census of effect-size practices in the past 5 volumes of American Speech-Language-Hearing Association journals was accomplished. Inclusion of effect size in quantitative research reports increased from 5 reports with effect size in 1990 to 1994 to 120 reports in 1999 to 2003. Nonetheless, effect size was reported less than 30% of the time when inferential statistics were used, and only half of those reports included an interpretation of effect size. This article presents case exemplars to illustrate the use and value of effect size and includes suggestions for interpreting effect size. Researchers are encouraged to routinely report effect size and to interpret effect size in a way that facilitates the application of research to practice.

Key Words: effect size, practical significance, clinical outcomes, evidence-based practice

The importance of research to the science of speech, language, and hearing is sometimes debated but never trivialized. Practitioners, teachers, and scientists alike depend on the body of research to further their individual goals. Clinicians seek to deliver the best services and to justify the effectiveness and efficiency of clinical services. Teachers seek knowledge to guide students in their discovery of truths or near truths. Scientists look for patterns of meaning as well as new ideas for improving the scope of knowledge in hearing, speech, and language. Speech-language pathologists (SLPs) and audiologists depend on research to advance the information base in speech, language, and hearing, and the information base is the heart and soul of the professions. Thus, quality of research is a matter of great importance to all.

SLPs and audiologists look to a variety of sources for information as well as points of dissemination. For example, they read peer-reviewed journals from the spheres of education, psychology, engineering, business, and others. However, the sciences (basic and applied) for communication disorders are largely defined by four peer-reviewed periodicals published by the American Speech-Language-Hearing Association (ASHA). Since 1991, ASHA has published the *American Journal of Audiology* (*AJA*), the *American Journal of Speech-Language Pathology* (*AJSLP*), *Language, Speech, and Hearing Services in Schools* (*LSHSS*), and the *Journal of Speech, Language, and Hearing Research* (*JSLHR*)—formerly, the *Journal of Speech and Hearing Research*. Although each of the four journals differs in purpose and scope, all four journals seek to advance knowledge in the scientific disciplines and the clinical professions. Just as the knowledge base in speech-language pathology and audiology grows, the knowledge base for research methods grows too. In other words, just as practitioners learn new and better ways to practice their craft, scientists learn new and better ways to design studies, analyze results, and formulate conclusions. The focus of this article concerns the latter two areas of research methods: (a) analyzing results and (b) formulating conclusions. Our purpose is to describe the current state of affairs in ASHA journals and to prescribe avenues for improving the science in communication disorders.

It seems to us that speech-language pathologists, audiologists, and others who engage in research and the practices concerning communication disorders possess an abundance of imagination and no shortage of novel ideas. However, the pursuit of knowledge requires sound methods as well. From time to time but perhaps not often enough, we see pleas to better the scientific methods in communication sciences and disorders. Three such expositions that come to mind are Max and Onghena (1999), Meline and Schmitt (1997), and Muma (1993). Max and Onghena reported some shortcomings in the statistical analysis of data, and they recommended appropriate remedies. Meline and Schmitt presented case studies of research to illustrate the uses of effect size, and Muma argued for the need for replication in the communication sciences and disorders scholarship. Each of these issues is of

1058-0360/04/1303-0202

current concern, and each issue will determine to a large degree the future of the sciences in speech, language, and hearing. This article revisits the issue of effect size, reports the current status for use of effect size in ASHA journals, and explains the importance of effect size with several illustrations.

The State of Affairs for Effect Size

About 7 years ago, Meline and Schmitt (1997) examined 411 research articles in ASHA journals and found that effect size was reported in only 5 of 411 articles. Meline and Schmitt used case studies from the speech, language, and hearing literature to argue for the inclusion and interpretation of effect-size statistics in research reports. Because 7 years had passed, we were interested in the change, if any, in reporting frequencies for effect-size statistics in ASHA journals. An increase in effect-size reports was expected, given the low frequency of effect-size reports in the earlier 5-year period (1990–1994) and the intervening publication of the fifth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001), which addressed the need for reporting effect size in research reports. To quantify the extent of change in effect-size reporting practices, all articles and reports published in ASHA journals from 1999 to 2003 were examined in the same fashion as Meline and Schmitt (1997) examined the 1990–1994 journals. Articles and reports were tabulated when t, F, χ^2, H, or another inferential statistic was designated as the omnibus statistic. By far, analysis of variance designs outnumbered the alternatives, so F was the primary statistic found in most studies. To ensure a high degree of reliability, the first author's judgments as to whether an effect size was reported were compared with results independently collected by two research assistants. The resulting index of reliability was a point-by-point agreement of 100%. A summary of results from the examination of 1999 to 2003 ASHA journals is displayed in Table 1.

In total, 433 articles were identified with one or more inferential statistics reported in the Results sections, a small increase from the 411 articles identified in the 1990–1994 period by Meline and Schmitt (1997). The data in Table 1 identify the number of articles that contained at least one effect-size metric. In the 1999–2003 period, 120 articles with effect-size metrics were identified, compared with 5 articles with effect-size metrics in the 1990–1994 period. The difference (120/5) was judged an important one given the large simple effect size ($\Delta = 115$ effect-size reports). Although the overall result suggested a significant change in effect-size reporting practices, an examination of statistics for each of the four journals was also revealing. Effect-size statistics were reported in 27.7% of the articles overall, but results for the individual journals varied widely and ranged from 72% (*LSHSS*) to 13% (*AJA*). Although many authors reported effect size, nearly half of the authors did not interpret their effect-size results. For example, 42 articles in the four ASHA journals included effect-size metrics in 2003, but only 23 (55%) interpreted the effect-size result. With a few exceptions, authors who interpreted effect-size results qualified the result as *small*, *medium*, or *large*—no more. To facilitate evidence-based practice, further explanations of effect-size results and their meaning for clinical practice are desirable (cf. Meline & Paradiso, 2003). For reporting results in journal articles, Bem (2004) recommended, "Whenever possible, state a result first and then give its statistical significance, but in no case should

Table 1. Effect-size reporting practices in journals of the American Speech-Language-Hearing Association, 1999–2003.

Journal volume	No. articles[a]	No. articles reporting effect size
American Journal of Speech-Language Pathology		
1999	8	2
2000	10	4
2001	8	4
2002	9	2
2003	25	11
American Journal of Audiology		
1999	1	0
2000	2	1
2001	4	0
2002	5	1
2003	3	0
Language, Speech, and Hearing Services in Schools		
1999	8	2
2000	7	5
2001	3	3
2002	3	3
2003	8	8
Journal of Speech, Language, and Hearing Research		
1999	56	9
2000	64	6
2001	57	7
2002	73	29
2003	79	23
Total	433	120 (27.7%)

[a] Number of articles with F, t, χ^2, H, or another inferential statistic as the primary test statistic.

you ever give the statistical test alone without interpreting it substantially" (p. 200). Clearly, effect-size reporting in ASHA journals has increased substantially from the early 1990s, but how does this result compare with effect-size reporting practices in other disciplines and other journals?

Effect-Size Reporting Practices in Non-ASHA Journals

Keselman et al. (1998) reviewed statistical practices in 17 contemporary education journals. They found effect-size metrics in 11% of articles with inferential methods, but the results varied widely by type of design and journal. Vacha-Haase, Nilsson, Reetz, Lance, and Thompson (2000) tabulated reporting practices for effect size in the 1990–1997 volumes of *Psychology and Aging* and the *Journal of Counseling Psychology*. They found effect-size report rates of 47% and 61%, respectively, in the 1997 volumes. Thompson and Snyder (1998) surveyed the 1996 volume of the *Journal of Counseling and Development* and identified 15 of 25 quantitative studies (60%) with at least one measure of effect size. Kirk (1996) surveyed four American Psychological Association (APA) journals, with the frequency of effect-size reports ranging from 12% to 77%. According to Kirk, the 1995 volume of the *Journal of Applied Psychology* contained reports of effect size in 77% of its articles. Finally, Paul and Plucker (2004) surveyed three gifted education journals published from 1995 to 2000, and they found a range of 24% to 34% for reporting effect size. Paul and Plucker found that only about half of the articles with effect size included some interpretation of the effect-size result. Overall, the reporting practices for effect size in non-ASHA journals varied widely from 11% to 77% depending on the specific journal and discipline. These percentages were very similar to those for current ASHA journals (13%–72%).

Editors and editorial boards—those of ASHA and others—clearly are recognizing the need for reports of effect size in their journals. In a 1997 editorial, Kevin R. Murphy announced the expectation of the *Journal of Applied Psychology* for authors to routinely include effect-size metrics unless there was a compelling reason not to do so. Other journals are requiring authors to include effect size when submitting articles or are strongly encouraging authors to do so. For example, *Personnel Psychology* includes the "report and discussion of effect size" in their article review checklist. The *Journal of Early Intervention* requires the inclusion and interpretation of effect-size metrics for all manuscripts. Other journals that require effect size are *Educational and Psychological Measurement*, the *Journal of Consulting Psychology*, the *Journal of Experimental Education*, the *Journal of Learning Disabilities*, *Language Learning*, *The Professional Educator*, and *Research in the Schools*.

In 1999, Wilkinson and the Task Force on Statistical Inference of APA's Board of Scientific Affairs recommended that authors "always present effect sizes for primary outcomes" (p. 599). In response to the task force recommendation, the fifth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001) states, "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" (p. 25). ASHA journals are among the many behavioral and social science periodicals that expect contributors to follow the style specified in the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001). ASHA editors Bahr (2001) and Peach (2003) affirmed the need for effect size in communication disorders research in their editorials. Additional arguments for the inclusion and benefits of effect size in research reports were presented by Cohen (1997), Kirk (2001), and Thompson (1999a, 1999b).

What Is Effect Size, and How Is It Different From Statistical Significance?

Effect size is a metric that estimates the size of a treatment effect. Unlike tests of statistical significance, effect-size metrics are unaffected by sample size. A large sample may enable an effect to reach statistical significance, but the effect may be trivial in importance. Alternatively, a small sample may fail to reach statistical significance although the result may be clinically important. To test statistical significance, researchers choose an acceptable level of chance for rejecting the null hypothesis before the experiment begins. By convention, an acceptable level of chance in the behavioral sciences is 5% (5 chances out of 100 for a Type I error). To measure effect size, researchers compute the difference between treatment means (treatment effect) and divide by the standard error. The calculation of effect size is accomplished at an experiment's conclusion. Effect size is actually a family of indices. There are many different metrics for effect size, and the choice of an effect-size metric depends on

the research design. For example, Robey (2004) described effect-size point and interval estimates for analysis of variance designs. Another resource for selecting an appropriate effect-size statistic is Kirk's (1996) tutorial. Kirk identified 40 effect-size metrics that have been proposed to help evaluate the importance of treatment effects.

Three Categories of Effect Size

There are three categories of effect-size metrics. The first is *simple* effect size, which is the raw difference between treatment means. For example, van Kleeck and Beckley-McCall (2002) recorded reading times for 5 mothers and their children. They reported average reading times (in minutes:seconds) for younger and older siblings. The averages were 1:27 and 5:09, respectively. In this case, the simple effect size was the difference between the means (effect size = 3:42). Simple effect size is useful when the unit of measurement is easily interpreted, but it has some inherent shortcomings. For one, it does not account for variability in the samples. In fact, the mothers in the van Kleeck and Beckley-McCall study varied from 0:57 to 1:50 when reading to their younger children and 3:02 to 8:17 when reading to their older children. Furthermore, simple effect size is not useful for comparing results from study to study.

These shortcomings are resolved by computing a *standardized* effect size. Standardized effect size accounts for variation between participants and provides a metric that is useful for comparing results between studies. For these reasons, standardized effect size is the metric of choice for meta-analytic (synthesis) studies (cf. Robey & Dalebout, 1998). For van Kleeck and Beckley-McCall's (2002) data, the standardized effect size would be computed by dividing the mean difference of 3:42 by the pooled standard deviation [square root of $(\sigma_1^2 + \sigma_2^2/2)$] of 2:28. The result is a standardized effect size (d = 1.5).

The third category of effect-size metrics is the effect-size *correlation*. Effect-size correlations are the correlations between the independent variable classification and the individual scores of the dependent variable. For example, Ingram and Morehead (2002) reanalyzed data collected from language-impaired and typically developing children in the 1970s. Their dependent variable was the occurrence of grammatical morphemes, and their independent variable was language status (impaired or typical). They reported an effect-size correlation (r^2 = .54) for progressive morphemes. The square of the correlation indicates the percentage of variance in progressive morphemes that was accounted for by membership in the independent variable groups (impaired and typical). In meta-analytic studies, rs are typically presented rather than r^2s. Otherwise, the choice of one or the other is dependent on the researcher's preference and the manner of interpreting results.

Interpreting Effect Size

Cohen (1988) provided some conventions for interpreting effect size. He suggested that an effect-size correlation of .5 was large, .3 was moderate, .1 was small, and anything smaller than .1 was trivial. For the standardized effect size d, Cohen suggested that .8 was large, .5 was moderate, .2 was small, and anything smaller than .2 was trivial. The problem with conventions of this sort is that they do not account for the unique properties of particular behavioral variables. For example, the expected effect sizes for adults with aphasia differs from that for children with language impairment. Meline and Schmitt (1997) reported typical standardized effect sizes for children with verb morphology as the dependent variable that ranged from 1.05 to 1.25. Robey (1998) reported an average standardized effect size equal to 0.61 for the treated and untreated recoveries of acute-stage aphasics. Thus, interpretations of effect size are best based on experience from earlier studies with the same dependent variable—a benchmark for prospective studies. If there is no experience to serve as a guide, the next best resource for interpretations of effect size may be Cohen's criteria.

Another informational resource for the interpretation of effect size is Becker's (2000) instructional module. Becker illustrated standardized effect sizes as percentiles of treated groups versus untreated groups. He also illustrated effect size as a percentage of nonoverlap for treated group scores versus untreated group scores. Percentiles and nonoverlap are alternative ways to think about effect-size results and may be useful for explaining results to consumers. In addition to Becker's tutorial, Meline and Paradiso (2003) presented some alternative ways to explain effect-size results along with case exemplars.

Illustrations of Effect Size in Communication Disorders Research

A Close Encounter of the Second Kind

Walker, Shinn, Cranford, Givens, and Holbert (2002) investigated the temporal processing abilities of college students with

reading disorders. They compared 9 students with reading disorders (experimental group) and 9 students without reading disorders (control group). Their data were percentages of correct scores from (a) a pitch pattern test, (b) a duration pattern test, and (c) a brief tone frequency difference limen (just noticeable difference) test. Walker et al. hypothesized a relationship between temporal processing skills and reading abilities. Based on statistical significance testing, they reported (a) no significant group effect for pitch patterns, (b) a significant group effect for duration patterns, and (c) no significant group effect for difference limens. Thus, they concluded that their participants with reading disorders showed a significant discrimination deficit for duration patterns but not for difference limens or pitch patterns.

A closer look at the Walker et al. (2002) results with the help of an effect-size evaluation provides meaningful information and further support for their conclusion. First, Walker et al. reported a nonsignificant effect between groups for the difference limen test. Their statistic is reported as $F(1, 16) = 4.257$, $p = .056$. Because of the small number of participants ($n = 18$), the F statistic failed to meet the .05 cutoff point. If Walker et al. had included 1 or 2 more participants, they might have successfully rejected the null hypothesis. The dilemma for researchers who experience a close encounter such as $p = .056$ is whether to accept the mathematical certainty of the .05 confidence level or to venture further analyses. Thompson (1999a) noted that "p values cannot be used as an effective vehicle for escaping disagreement and confrontation regarding our subjective judgments of the worth of our results" (p. 168). Clearly, behavioral researchers cannot respond to demands for practical results and ignore human values at the same time. Statistical significance tests have nothing to do with practical significance. On the other hand, effect-size metrics have everything to do with practical significance. Thus, effect-size metrics are an important avenue for bridging the gap between research and practice.

Effect-size statistics are appropriate as follow-ups to inferential statistics, such as F and t, as well as distribution-free tests, such as the Kruskal–Wallis H; alternatively, effect-size statistics may stand alone. Effect-size metrics can be computed for differences between means that do not otherwise achieve statistical significance. In the case of the Walker et al. (2002) result for difference limens between groups ($p = .056$), the most disparate means for the two groups were 14.6 Hz (control) and 30.6 Hz (reading disorder). In this instance, the mean difference was 16 Hz. A standardized effect size would be computed as 16 Hz divided by [square root of $(12.7^2 + 6.2^2)/2$] = 1.60. Thus, $d = 1.60$ and might be interpreted as a large treatment effect. Walker et al.'s contention that adults with reading disorders exhibit problems with the discrimination and perception of auditory stimuli is further supported by evaluating effect sizes.

Walker et al. (2002) also reported no statistical significance between groups for pitch patterns. However, they reported means of 90.9% for the control group and 72.8% for the experimental group. Although not statistically significant, the difference between the means was 18.1%. A standardized effect size would be computed as 18.1% divided by [square root of $(9.7^2 + 8.5^2)/2$] = 1.98. Thus, $d = 1.98$ and might be interpreted as a large treatment effect. Again, the result adds practical support for the Walker et al. hypothesis.

Back to the Future With Grammatical Morphemes

Ingram and Morehead (2002) reanalyzed data that they first reported in 1973. For the reanalysis, they examined the occurrence of five grammatical morphemes spoken by 6 language-impaired and 6 typically developing children. The analysis included progressive, plural, possessive, third-person singular, and regular past tense morphemes. The only statistical significance was between groups for the progressive morpheme in all contexts as well as in obligatory contexts (i.e., those contexts that require the correct form as specified by the target language). The language-impaired group produced more progressive morphemes in all contexts (7.2% vs. 1.8%) and in obligatory contexts (69% vs. 27.8%). The language-impaired group also evidenced an advantage for possessive morphemes in obligatory contexts (42 % vs. 14%), but the difference was not statistically significant ($p > .05$). Nevertheless, we calculated a standardized effect size (d = 28%) divided by [square root of $(44^2 + 16.9^2)/2$] = 0.84. The observed effect size ($d = 0.84$) appears to be a moderate treatment effect based on past experiences with effect size, grammatical morphemes, and language-impaired children. Although not statistically significant, the language-impaired children's advantage with possessive morphemes was a practical effect that may have meaning in everyday affairs such as clinical practice.

Conclusion

To advance the science of speech, language, and hearing in the journals and to improve the

practical significance of research for clinicians, it is imperative that methods and analyses are improved in research endeavors. Researchers should routinely (a) include estimates of effect size and (b) interpret effect-size metrics within the context of their experiment. These outcomes will strengthen the conclusion validity in research reports, help to bridge the research to practice divide, and benefit the scientific base for audiology and speech-language pathology.

References

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

Bahr, R. H. (2001). From the editor. *Language, Speech, and Hearing Services in Schools, 32*, 3.

Becker, L. A. (2000). *Effect size (ES)*. Retrieved April 14, 2002, from University of Colorado at Colorado Springs, Department of Psychology Web site: http://web.uccs.edu/lbecker/Psyc590/es.htm

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A career guide* (2nd ed., pp. 185–219). Washington, DC: American Psychological Association.

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

Cohen, J. (1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21–35). Mahwah, NJ: Erlbaum.

Ingram, D., & Morehead, D. (2002). Morehead & Ingram (1973) revisited. *Journal of Speech, Language, and Hearing Research, 45*, 559–563.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

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*, 213–218.

Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 42*, 261–270.

Meline, T., & Paradiso, T. (2003). Evidence-based practice in schools: Evaluating research and reducing barriers. *Language, Speech, and Hearing Services in Schools, 34*, 273–283.

Meline, T., & Schmitt, J. F. (1997). Case studies for evaluating significance in group designs. *American Journal of Speech-Language Pathology, 6*(1), 33–41.

Muma, J. (1993). The need for replication. *Journal of Speech and Hearing Research, 36*, 927–930.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3–5.

Paul, K. M., & Plucker, J. A. (2004). Two steps forward and one step back: Effect size reporting in gifted education research from 1995–2000. *Roeper Review: A Journal on Gifted Education, 26*, 68–72.

Peach, R. K. (2003). From the editor. *American Journal of Speech-Language Pathology, 12*, 258.

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research, 41*, 172–187.

Robey, R. R. (2004). *Reporting point and interval estimates of effect size for fixed between-effect analysis of variance*. Manuscript submitted for publication.

Robey, R. R., & Dalebout, S. D. (1998). A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research, 41*, 1227–1241.

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

Thompson, B. (1999b). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*, 191–196.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *Journal of Counseling and Development* research articles. *Journal of Counseling and Development, 76*, 436–441.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory and Psychology, 10*, 413–425.

van Kleeck, A., & Beckley-McCall, A. (2002). A comparison of mothers' individual and simultaneous book sharing with preschool siblings: An exploratory study of five families. *American Journal of Speech-Language Pathology, 11*, 175–189.

Walker, M. M., Shinn, J. B., Cranford, J. L., Givens, G. D., & Holbert, D. (2002). Auditory temporal processing performance of young adults with reading disorders. *Journal of Speech, Language, and Hearing Research, 45*, 598–605.

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

Received May 23, 2002
Accepted June 4, 2004
DOI: 10.1044/1058-0360(2004/021)

Contact author: Timothy Meline, PhD, The University of Texas—Pan American, Department of Communication Sciences and Disorders, 1201 West University Drive, Edinburg, TX 78541-2999. E-mail: TM2776@AOL.COM