Power and Effect Size Measures: A Census of Articles Published from 2009 -2012 in the Journal of Speech, Language, and Hearing Research

Manish K. Rami, PhD
Associate Professor
Communication Sciences and Disorders
University of North Dakota
Grand Forks
ND 58202-8040

Abstract

The purposes of this study were to report the use of power and effect sizes in articles in the Journal of Speech, Language, and Hearing Research (JSLHR). A census of all articles published between the years 2009 - 2012 (n = 436) in JSLHR was conducted to gather data about report of power and effect size. Over 97% (426/436) of the articles did not report power and over 42% (187/436) did not report any effect size measure. Articles not reporting power in each year in this census range from 96.5% to 98.9% and those not reporting any effect size ranges from 33% to 50% over a period of four years. The discipline would benefit from embracing higher standards and encouraging authors to report the measures of power and effect size in their articles.

Keywords: Power, Effect Size, Speech-Language Pathology, Audiology, Communication Sciences and Disorders

This study briefly explains the statistical concepts of power and effect size, reports the results of a census of all the articles published during the year 2009 -2012 in the Journal of Speech, Language, and Hearing Research (Journal) and highlights the need for a consistent and prevalent use of these concepts. Some resources for learning about and appropriate use of power and effect size are provided.

1. Power

In determining the state of the null hypothesis, a researcher typically expects the experimental data (data) produced to be in agreement with the real world facts (facts). If the data are indeed in agreement with the facts, then the researcher is more likely to make one of the two correct decisions. These two correct decisions are to either 1) to retain the null hypothesis when it is true or 2) to reject the null hypothesis when it is false. However, the risk that the data do not agree with the facts is always present. In cases wherein the risk of disagreement between the data and the facts is high, a researcher is more likely to make one of two incorrect decisions. These incorrect decisions are to either 1) reject the null hypothesis when it is true and commit the type I error or 2) accept the null hypothesis when it is false and commit the type II error.

1.1 Risks

The risks that lead to committing these errors are many. One way to minimize the risk of committing a type I error is to set a low probability of rejecting the null hypothesis when it is true. This probability (the alpha) is commonly and arbitrarily set at .05. However, decreasing the alpha too much increases the probability of making a type II error (Cohen 1992a; Keppel, 1991; Stevens, 1996). The way to control type II error without increasing the risk of committing type I error is to increase the power (or the sensitivity) of the experiment. Increasing the power improves the chances of finding the data that aligns with the facts. Though researchers have accepted the common practice of controlling type I error by setting a low alpha level, there may not be an adequate effort to control the type II error by increasing the power of the experiments. (Brewer, 1972; Cohen, 1962; Cohen 1992a; Jones, Gebski, Onslow, & Packman, 2002). Four times as much risk in committing type II error is common (Cohen, 1962). That is, the ratio of type II error to type I error is 4:1. It is clear that researchers give more importance to type I errors than to type II errors. However, a power analysis is deemed essential to assess the suitability of an experimental design and to quantify the risk of committing a type II error. (Winer, Brown, & Michels, 1991).

1.2 Reports

Power reports in the Journal and within at least one specific area in the profession are not common. Young (1993) provided an anecdotal account of lack of report of power in the Journal. A survey investigating the frequency of report of power in the Journal was not very encouraging (Rami, 2010a). The findings in this survey indicated that 98.9% of the articles did not report power. Specifically within the field of stuttering, Jones, Gebski, Onslow, & Packman (2002) after surveying 26 studies in stuttering, reported that authors do not seem to be concerned about power in their studies. Researchers seem, perhaps inadvertently so, to disregard any risk of committing type II error that might be present in their experiments. In the absence of the knowledge of power in an experiment, it is difficult to estimate the risk of having accepted a null hypothesis when in fact the null hypothesis was false or commit a type II error. Unfortunately, such lack of report of power in experimental research in the Journal is in line with several reports from diverse fields. Reports of underpowered studies are documented in abnormal-social psychology (Cohen, 1962), education research (Brewer, 1972), criminal justice (Brown, 1989), orthopedics (Freedman, Back, & Bernstein, 2001), behavioral ecology and animal behavior (Jennions, & Møller, 2003), and operations management (Verma, & Goodale, 1995).It is unknown if this practice has changed since Young (1993) and Jones, Gebski, Onslow, & packman (2002) or if authors of articles in the Journal are reporting power as a matter of routine practice.

2. Effect Size

Researchers plan their experiments carefully to isolate the effects of an independent variable (IV) on the dependent variable (DV). Finding statistical significance in itself, however, does not provide a researcher with any estimate of the association between the IV and the DV or the extent of control of the DV by the IV. The various indices that quantify the association or the extent of control that the IV exercises on the DV are known as effect sizes. Some examples of effect size (ES) measures are R squared, Cohen's d, Eta Squared, and Omega squared. The knowledge of ES of an IV –especially if the IV is a treatment, allows a clinician to effectively manipulate the IV, say in the course of intervention. This knowledge of the ES also allows the clinician to reliably predict the DV or the treatment outcome. Such knowledge allows clinicians to arrive at an informed and reliable prognosis. Hence, knowing the ES of an IV (say a type of treatment) is clinically very valuable information. Recommendation of the use of ES in research is widely made (Cohen, 1992b; Cohen, 1994; Glass, 1976). An additional advantage of reported ESs is that it can help future experimenters calculate power prospectively for their research work (Rosnow, & Rosenthal, 1983). It should also be noted that ES measures are comparable across experiments (Glass, 1976). This property of ES makes it possible to calculate the average effect size of an IV of interest from multiple studies. Such studies, known as meta-analyses, are explicitly conducted with an intention to better estimate the effect of an IV on the DV (Glass, 1976; Glass, McGaw, & Smith, 1981). Reporting ES is encouraged and found useful even in the absence of statistically significant results. (Zumbo, & Hubley, 1998) as such reports can be included in a meta-analysis. Thus, there is a further value to reporting ES in research studies. The use of ES or average ES in research as well as clinical decision-making is indispensable. However, the percentage of articles in any given year between 1999 and 2003 in the Journal that report ES varied from 9.37% to 39.72% (Meline & Wang, 2004.) The result from a survey of 91 articles in six volumes of the Journal in the year 2009 found a slight improvement in the report of ES, which was about 57.1% of the articles (Rami, 2010b).

3. The APA Style

The Journal has adopted use of the APA style (ASHA, 2012) and expects all authors to adhere to this style. The Publication Manual of the American Psychological Association (the Manual) (APA, 2001) which describes the APA style, distinctly addresses the issue of reporting power and effect sizes in its discussion on developing the results section. The Manual states that authors should report power and ES. Additionally, in the section on reporting statistics, the Manual emphasizes inclusion of adequate information such as the actual calculated values (of say χ^2, *t, F* etc.), the various degrees of freedom, and the *p* values. These values can help other researchers calculate ES when conducting a meta-analysis if ES is not reported (Hunter, & Schmidt, 2004). In essence, the Manual urges authors to report an adequate amount of information to permit the readers to understand not only the authors' interpretation of the statistics but also entertain other possible explanations. Others have made similar suggestions as well (Cohen, 1988, Keren, & Lewis, 1993). The advantage of clearly reporting all the relevant statistics is that such reporting facilitates the evaluation of each study for the sake of meta-analyses.

Therefore, not only is it important to report all the statistics for the sake of readers' individual interpretation of the results but it also facilitates further evaluation of the variables involved in a study via meta-analysis. There is currently no information on the consistency in formats when reporting statistics in the Journal.

The purpose of this census was to examine the report of power and ESs in all articles in the Journal post-2008.

4. Method and Results

All articles from each of the six volumes of the Journal for the years 2009 -2012 or 24 volumes total were examined. The articles published in the last four years were selected as they are deemed to reflect current practice in the discipline. All articles reporting at least one inferential test were selected. Letters to the editor and review articles that did not report any statistics were excluded from this census. Any report of power and ES in an article, even if it was for one or some test results, was deemed as an acceptable report (n = 436 articles). Brief memos were created for those articles that failed to report the actual values of statistical test results or used a non-standard format for reporting results.

4.1 Qualitative Analysis

Qualitatively, varieties of differences were found. From at least one article reporting the measures of power and ES for every test conducted in the study, to some merely using a non-standard format, to at least one reporting results of *t* tests, χ^2 tests, and *F* tests, but not reporting any calculated values or any degrees of freedom. The last case, would fail to contribute to a meta-analysis.

4.2 Quantitative Analysis

Quantitatively, the percentages of articles reporting power and ES were calculated for each year. It was found that 97.7% of the articles (426/436) did not report power and 42.88% of the articles (187/436) did not report an ES. Figures 1 and 2 show the volume-by-volume variations in the number of articles reporting and not reporting power and ES for the years 2009- 2012.

5. Discussion

5.1 Power

Only 2.3% of the authors of all the articles reviewed (10/436) reported power values. Such lack of reporting of power in experimental research in the Journal is in line with several other reports from diverse fields (Cohen, 1962; Brewer, 1972; Freedman, Back, & Bernstein, 2001; Jennions, & Møller, 2003; Verma, & Goodale, 1995). Experiments with low power fail to detect the presence or absence of differences or relationships of interest, decrease the reliability of the findings, and waste effort and resources used for research.

The reasons for not reporting power values are unknown.(Jones, Gebski, Onslow, & Packman (2002) suggested that perhaps a majority of authors disregard the risk of committing type II error that might be present in their experiments. This could be the case if the authors thought that reporting power did not affect the interpretation of their data. However, that is not the case (see example below.) The other possibility, although doubtful, could be inadequate statistical training or an over emphasis on controlling type I error. While the prior is a conjecture, the brief discussion below could be helpful. If the latter were the case, a change in the amount of risk in committing the type II error as compared to the type I error would be necessary.

5.1.1 Ways to Increase Power

There are some methodological ways to increase power. These include choosing the most sensitive experimental design or choosing specific statistical procedures (Keppel, 1991). However, the three primary determinants of power are the size of the sample, the size of the effect under investigation, and the significance level alpha. The size of the sample is routinely suggested as a way to increase power in an experiment. Increasing the number of observations/participants in an experiment should be considered whenever possible. However, this approach is often difficult. Another way to increase power is by choosing variables, as possible, that could produce large effects. Very small treatment effects might not be of clinical use. Finally, the way that most researchers seem to overlook, perhaps even ignore, is relaxing significance level in favor of increasing power. This suggestion, though it might seem sacrilegious, is not new. Several prominent statisticians have made this suggestion over a period of half a century (Cohen, 1962, 1988, Keppel, 1991, Neyman, 1957, Overall, 1969).

5.1.2 Ways to Calculate Power

There are three way to deal with calculation of power. One, experimenters can and ideally should conduct a power analysis prior to conducting a study. In order to conduct a power analysis, one needs to estimate the likely ES. Effects size estimates can be obtained either by examining existing research or by assuming moderate to large ESs. Another and a better but more expensive approach to quantify power is to first conduct a pilot study to quantify an ES and then use those estimates in a power analysis. Finally, the least that a researcher can do is to report the power estimate as calculated by statistical software. Power analysis can be conducted using one of the several soft wares available, for example, SPSS and Gpower. Several articles explaining the importance and appropriate use of power analysis are available. (For example, see Cohen, 1988; Cohen, 1992a; Lenith, 2001; Mayr, Erdfeler, Buchner, & Faul, 2007; Zumbo, & Hubley, 1998.) For a short of list of resources on power, see Appendix. A.

6. Effect Size

Over 42% of the authors of all the articles reviewed (n = 436) did not report ES values. This number compares favorably with Meline and Wang (2004) who reported that about 60% of the authors of articles in the Journal did not report ES between the years 1999 – 2003. It is likely that perhaps the authors of articles in the Journal do not deem ES as an important measure to calculate and report. Effect size measures are good estimators of the amount of control an IV has over DV. Without the knowledge of the ES, it is difficult to estimate either the experimental or the clinical value of manipulating the IV to produce a desired change in the DV. The ES information is crucially important if the IV happens to be a treatment. Therefore, it is imperative that ES measure be routinely reported in every experimental study published regardless of the statistical significance found in the inferential tests (Zumbo, & Hubley, 1998).

Depending on the design of a researcher's experiment, an appropriate measure of ES should ideally be calculated and reported along with a note about the interpretation of the ES measure used (see Tatsuoka, 1992). At the minimum, for example, a study correlating variables could report the R^2 or the ε^2, while a study comparing means could report the Cohen's d. Commonly these days, the statistical software used to analyze the data have an option to calculate various ES measures. An appropriate option can be selected to calculate and report ES along with a guide to its interpretation. For a short of list of resources on ES, see Appendix. B.

6.1 An Example

In order to better understand the consequences of not knowing the power or the ES in an experiment let us consider an example. Consider an experiment designed to explore the effects of a particular treatment (T) on some speech or language behavior (B). Let this be a pre-treatment, post-treatment true experimental design with an experimental and a control group. Let the participants be all randomly selected from the population of interest to increase external validity. These participants are then randomly assigned to one of the two groups to assure internal validity. The experiment is then run beginning with pre-treatment measurement of B in both the groups. Next, all the participants in the experimental group receive T under investigation while the control group participants do not receive any T. Finally, after the treatment is completed in the experimental group, the B is measured again in both the groups providing the post-treatment scores. While the null hypothesis in this experiment would assume no differences in post-treatment measures between the two groups, the primary goal in such an experiment is to examine any differences in post-treatment measures of the DV between the experimental and the control groups. The crucial question from the clinical perspective is not just whether T will bring about change in B or not but also how much change could T bring in B. The prior question can be answered by conducting significance tests and the latter by calculating ES. If the participants in the experimental group perform better on B than those in the control group, then T shows merit in its use. In the alternative, if either T fails to improve or decreases B in the participants in the experimental group, then the outcome is noted but it does not merit clinical use.

One of the two types of errors to avoid in the above instance would be to conclude that T is useful when it is not really so or reject the null when it is true (type I error.) The chance of making type I error varies from 0 – 1. However, it is commonly and arbitrarily set at 0.5. Type I error might result in the use of T with no resulting change in B or any benefit for the patients and a waste of valuable resources such as time, effort, and money. The other error to avoid is to find T useless when in reality it is useful or accept the null when it is false (type II error.) The chance of making a type II error also varies from 0 -1. If type II error is controlled, it is set at .2.

This is known as beta and it translates to power of .8, which is deemed an acceptable level of power (Cohen, 1988). A type II error would result in a missed opportunity to establish the usefulness of the treatment (T) investigated and to help the patients. While every effort should be made to avoid both of these errors, power specifically addresses the type II error. If the power in such an experiment is low, then the likelihood of committing a type II error increases. A low powered study not only depletes valuable resources but it also obstructs researchers and practitioners from discovering and employing useful treatments for the benefit of patients. One way to avoid making such an error is to conduct a power analysis before the experiment is run to ensure that the experiment will not result in a failed effort. It behooves us to demonstrate that type II errors are controlled as much as possible and not likely to occur in our experiments. Another, an easier and uncomplicated approach, is to report the power as calculated by most statistical soft wares during the inferential analysis. As consumers of research products, one should keep in mind that a low powered experiment does nothing to help us develop an informed opinion about the usefulness of a treatment under examination.

Concerning the example, let us further assume that our experiment has adequate power and that it is successful in finding statistically significant differences between the post-treatment measures of the two groups. The presence of a statistical significance does not provide one with any quantification of the extent of control T has on B, however. This means that a practitioner has no means of predicting the magnitude of the effect a treatment. Without knowing this ES, it is not possible to determine the usefulness or the clinical benefit of the treatment. In experiments with a large enough sample and high power, very small effects might be detected and found to be statistically significant. In such cases, the use of the treatment found to be statistically significant might not bring any major or quantifiable change in the targeted behavior, as the ES is small. Thus, a statistically significant effect does not always translate into clinical usefulness of the treatment. As consumers, one should keep in mind that the ES, though invariably influenced by substantive issues, ought to be at least medium if not large. If the ES is small, changes in B, if there are any, might not be measurable.

It is hoped that this brief discussion of the statistical concepts of power and effect size highlights the risks of ignoring these measure. Perhaps a suitable adjustment of the risk of committing the type I and type II error, perhaps in the ratio of 1:1 ought to be considered. It is also hoped that the Journal recommends and colleagues adopt the practice of reporting power, ES, and the results of any inferential tests in sufficient detail.

References

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

American Speech-Language-Hearing Association (ASHA). (2012). *Style Manual.* Retrieved July 30, 2012 from http://jslhr.asha.org/misc/ifora.dtl#style

Brewer, J. K. (1972). On the power of statistical tests in the American Education Research Journal. *American Education Research Journal, 9*(3), 391-401.

Brown, S. E. (1989). Statistical power and criminal justice research. *Journal of Criminal Justice, 17*(2), 115, 122.

Cohen, J. (1962). The statistical power of abnormal –social psychological research: A review. *Journal of abnormal and social psychology, 65*(3), 145-153.

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

Cohen, J. (1992a). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J. C. (1992b). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Cohen, J. C. (1994). The Earth is round. *American Psychologist, 49*(12), 997-1003.

Freedman, K. B., Back, S., & Bernstein, J. (2001). Sample size and statistical power of randomized controlled trials in orthopedics. *The Journal of Bone and Joint Surgery, 83-B*, 397-402.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research, 5*(10), 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis is social research.* Beverly Hills, CA: Sage Publications.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings.* Thousand Oaks, CA: Sage Publications, Inc.

Jennions, M. D., & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14*(3), 438-445.

Jones, M., Gebski, V., Onslow, M., & Packman, A. (2002). Statistical power in stuttering research: A tutorial. *Journal of Speech, Language, and Hearing Research, 45*(2) 243-255.

Keppel, G. (1991).*Design and analysis: a researcher's handbook.* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Keren, G., & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: methodological issues.* Hillsdale, NJ: Lawrence Erlbaum Associates

Lenith, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55*(3), 187-193.

Mayr, S., Erdfeler, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology, 3*(2), 51-59.

Meline, T., & Wang, B. (2004). Effect-Size Reporting Practices in AJSLP and Other ASHA Journals, 1999–2003. *American Journal of Speech-Language Pathology, 13*(3) 202-207.

Neyman, J. (1957). Inductive Behavior as a Basic Concept of Philosophy of Science. *International Statistical Review, 25(*1), 7-22.

Rami, M. K. (2010a). *Report of power in articles in JSLHR, 2009: A survey.* Poster presented at the American Speech-Language-Hearing Association Annual convention, Philadelphia, PA. *Asha Leader. Available at:* http://arts-sciences.und.edu/communication-sciences-disorders/_files/docs/rami2010a.pdf

Rami, M. K. (2010b). *Effect size reports in articles in JSLHR, 2009: A survey.* Poster presented at the American Speech-Language-Hearing Association Annual convention, Philadelphia, PA. *Asha Leader. Available at:* http://arts-sciences.und.edu/communication-sciences-disorders/_files/docs/rami2010b.pdf

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276-1284.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tatsuoka, M. (1992). Effect Size. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in behavioral sciences: methodological issues.* (pp.461-479). Hillsdale, NJ: Lawrence Erlbaum Associates.

Verma, R., & Goodale, J. C. (1995). Statistical power in operations management research. *Journal of Operations Management, 13,* 139-152.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). NY: McGraw Hill.

Young, M. A. (1993). Supplementing tests of statistical significance: variation accounted for. *Journal of Speech and Hearing Research, 36,* (4) 644-656.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician, 47*(2), 385-388.

Appendix A Resources for the use of Power

For a brief introduction to power, see Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155-159.

For an introduction to power and of ways to calculate power (examples from stuttering), see Jones, M., Gebski, V., Onslow, M., & Packman, A. (2002). Statistical power in stuttering research: A tutorial. *Journal of Speech, Language, and Hearing Research, 45,* 243-255.

For a comprehensive review of power and its use, see Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

For a tutorial on Gpower, a free power analysis software, see Mayr, S., Erdfeler, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology, 3*(2), 51-59.

Appendix B Resources for the use of Effect Size

For a tutorial on eta squared and omega squared, see Young, M. A. (1993). Supplementing tests of statistical significance: variation accounted for. *Journal of Speech and Hearing Research, 36,* (4) 644-656.

For a comparison of eta squared and omega squared, see Maxwell, S. E., Camp, C. J., &Avery, R. D. (1981). Measures of strength of association: a comparative examination. *Journal of Applied Psychology, 66*(5), 525-534.

For techniques of calculating ES specific to a design, see Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

For an introduction to ES and its valuation, see chapter 4 in Keppel, G. (1991). *Design and analysis: a researcher's handbook.* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Figure 1: Number of Articles (n= 436) in Volumes 1 -6 of the Journal for the Years 2009-2012 Reporting and not-Reporting Power

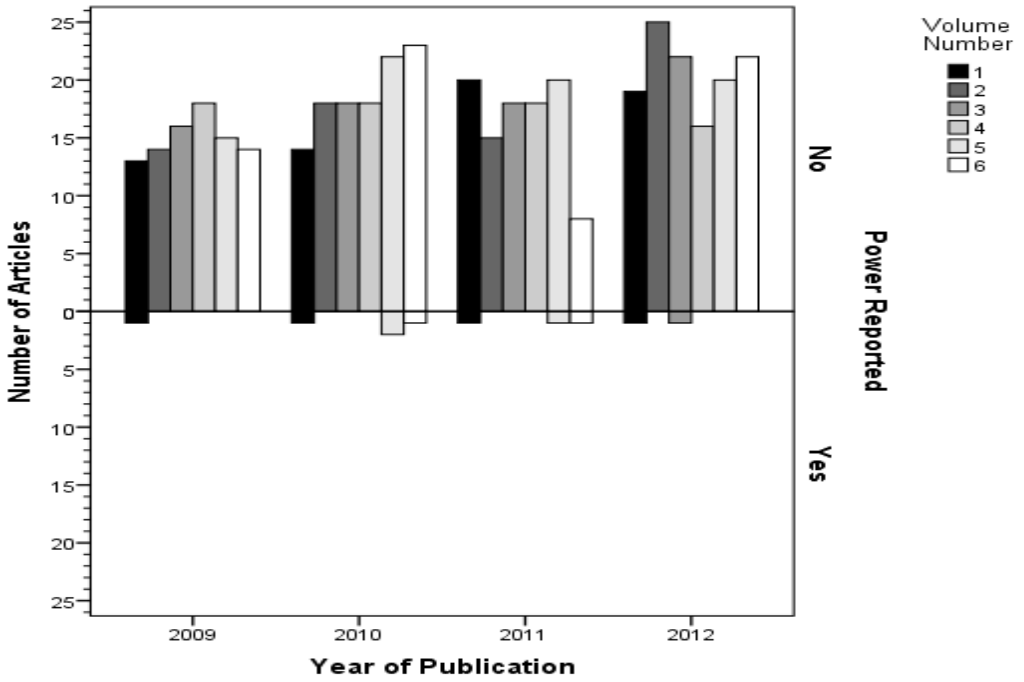

Figure2: Number of Articles (n= 436) in Volumes 1 -6 of the Journal for the Years 2009-2012 Reporting and not-Reporting Effect Size