International Journal of Humanities and Social Science Research
ISSN: 2455-2070
www.socialresearchjournals.com
Volume 2; Issue 1; January 2016; Page No. 22-25

Examining the utility of effect size calculations and percentage of non-overlapping data in single-case design research

Elias Clinton

College of Education and Behavioral Sciences, Black Hills State University Spearfish, South Dakota 57799, United States of America

Abstract
The use of effect size measures to bolster research findings has increased over the past 15 years. In general, the inclusion of effect sizes is encouraged by many in the scientific community to increase the magnitude of research findings. However, the inclusion of effect size calculations has become an area of controversy amongst scholars who utilize single-case research designs. This paper discusses the overall construct of effect sizes, the controversy of effect size metrics as it relates to single-case research, and an examination of the utility of one commonly used effect size metric in single-case designs: the percentage of non-overlapping data.

Keywords: single-case design research, single-subject design research, research methodology, effect size, non-parametric methods

1. Introduction

In 1999, the American Psychological Association (APA) Task Force on Statistical Inference encouraged researchers to include effect sizes for primary outcomes of empirical studies (Kelley & Preacher, 2012; Nakagawa & Cuthill, 2007; Wilkinson & American Psychological Association, 1999) [1, 2, 3]. Over the course of the past 15 years, the push for researchers to include effect sizes in their study summaries has increased (Nakagawa & Cuthill, 2007) [2]. For example, the current (sixth) edition of the APA publication manual includes a statement on effect sizes that reads: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section" (p. 34). The purpose of this paper is to provide a definition of effect size, as well as a discussion of effect size measures related to both group design and single-case design research.

2. Definition of effect size

An effect size quantifies the size of the difference between two groups (Kelley & Preacher, 2012; Nakagawa & Cuthill, 2007; Tuckman & Harper, 2012) [1, 2, 4]. Tuckman and Harper (2012) reported "effect sizes are represented by the ratio of the size of a difference between the means of two distributions to their average standard deviations (or the larger of the two)." Effect size estimates allow researchers to determine whether a treatment effect is small, medium or large. For research that involves correlation, effect size estimates allow researchers to determine the strength of the relationship, not just whether the relationship is statistically significant (Huck, 2012) [5]. The empirical literature outlines several effect size measures such as: standard mean difference, regression coefficient, Cohen's d, and correlation coefficient (Nakagawa & Cuthill, 2007) [2].

3. Effect size in group design research

Historically, the predominant statistical approach for group design research was null hypothesis significance testing (Kelley & Preacher, 2012; Nakagawa & Cuthill, 2007) [1, 2]. The statistical term *significance* indicates a level of confidence regarding the difference in groups' performance, or the existence of a relationship between variables (Shaver, 1993)[6]. Researchers conduct significance testing by determining an alpha level (i.e., an acceptable error rate), collecting data, calculating a specific statistic, and comparing that statistic to the determined critical value (Shaver, 1993) [6]. If the statistic is higher than the critical value, the finding is considered statistically significant and the researchers can reject the null hypothesis (i.e., the statement that there is no difference or relationship between the variables); therefore, it can be inferred that the probability is small that the difference or relationship happened by chance or sampling error alone (Shaver, 1993) [6]. If the statistic is lower than the critical value, than the finding is not significant and the researchers fail to reject the null hypothesis; therefore, the probability is high (e.g., <.05) that the difference or relationship happened by chance or sampling error alone (Shaver, 1993) [6].

Scholars have indicated that significance testing alone may not be an adequate or sufficient measure to determine treatment effectiveness because it does not provide specific information such as: the magnitude of an effect of interest or the precision of the estimate of the magnitude of an effect of interest (Huck, 2012; Nakagawa & Cuthill, 2007) [5, 2]. Some have even contended that null hypothesis testing should be abandoned altogether in favor of effect sizes and confidence intervals (Kelley & Preacher, 2012; Schmidt, 1996) [1].

4. Effect size in single-case design research

Calculating and including effect sizes in research reports has also become a topic in single case design research (SCD). The standard method of data examination in SCD is not statistical measures, but the use of visual analysis (Gast & Spriggs, 2014) [7]. Visual analysis is used in SCD to: determine stability of data, determine patterns in responding, compare data across phases to determine if manipulation of the independent

variable was associated with changes in the dependent variable, and to determine whether multiple demonstrations of effect were demonstrated (Parsonson & Baer, 1978; Kratochwill *et al*., 2010; Horner *et al*., 2005) [8, 9, 10]. In sum, visual analysis is used in SCD to determine the existence, or lack thereof, of a functional relation (i.e., a causal relationship between an independent and dependent variable).

Despite the traditional and logical use of visual analysis, scholars are increasingly investigating the application of calculating effect sizes within SCD to determine the magnitude of treatment effect (Wolery, Gast, & Ledford, 2014) [11]. However, there is currently no consensus on the logistics of using effect sizes in SCD. One specific issue is that single case research does not typically target the magnitude of the treatment effect, but historically demonstrates replication of the effects of treatment data patterns such as changes in level, trend, variability, and immediacy of effect in order to determine a functional relation (Gast & Ledford, 2014)[12]. Several methods of calculating effect sizes in SCD have been proposed in the literature such as: regression models (Manolov, Solanas, Sierra, & Evans, 2011) [13], and hierarchical linear modeling (Campbell, 2013)[14]. However, disadvantages have been documented for each of the aforementioned methods, and no consensus exists regarding which metric should be used to meaningfully determine effect size in SCD (Campbell, 2004; Haardörfer, 2010; Wolery, Busick, Reichow, & Barton 2010) [15, 16, 17].

5. Percent of Non-Overlapping Data

One method for determining effect sizes for SCD is Percent of Non-overlapping Data (PND) (Scruggs & Mastropieri, 1998) [18]. PND is a commonly used metric used to measure treatment effectiveness in SCD. While PND can be an effective tool to augment analysis of SCD data, there are documented limitations of this method as a metric for calculating ES. PND is a non-parametric overlap method used to determine the magnitude of intervention effectiveness (Scruggs & Mastropieri, 1998) [18]. The underlying assumption of the method is that a higher PND suggests the intervention was highly effective for facilitating desired changes in behavior (Scruggs & Mastropieri, 1998) [18]. Scruggs and Mastropieri (1998) [18] outlined specific criteria for interpreting PND scores: PND ranges from 0 - 100%, PND less than 50% reflects ineffective treatment, PND 50% - 70% indicates minimal effectiveness, PND 70% - 90% indicates moderate effectiveness, PND > 90% indicates a highly effective treatment. PND cannot be calculated when a zero quantity is in the baseline of decreasing behavior interventions or if the maximum possible quantity occurred during baseline of increasing behavior studies (Scruggs & Mastropieri 1998) [18]. Further, the use of PND is not recommended for data sets involving analysis of trend. Given the aforementioned limitation, the use of PND to measure treatment effectiveness would likely be inappropriate for any study involving initial skill acquisition as the dependent measure. For example, if a researcher were investigating the effects of a particular instructional package on the acquisition of completing complex math algorithms (i.e., long division) by a student with a disability, then the data would likely show a gradual, accelerating trend over time. If PND was calculated for the aforementioned scenario, a low treatment effect would be indicated, yet this would be erroneous as the student was making adequate gradual progress over an extended period of time.

Wolery *et al*. (2010) [17] compared overlap methods, including PND, to the visual analysis judgment of the authors for 160 data sets from the Journal of Applied Behavior Analysis. The authors reported the use of overlap methods as having unacceptably high levels of error for calculating ES. The authors surmised that overlap methods fail to detect all characteristics of time series data such as variability/trend. The authors also reported that overlap methods do not account for the replication logic inherent to SCD. The authors concluded that overlap methods are a wholly inappropriate metric for quantitatively analyzing treatment effects in single-case data and suggested that the methods be ultimately abandoned.

Allison and Gorman (1993) [19] reviewed methods for calculating ES for SCD. The authors reported that PND was an unsuitable metric for calculating ES for several reasons. The authors stated that PND is not a valid measure when single-case data sets contain outliers. See *Figure 1.* for an example provided by the authors in which an intervention has a positive effect, yet the presence of an outlier yields a PND score of 0.

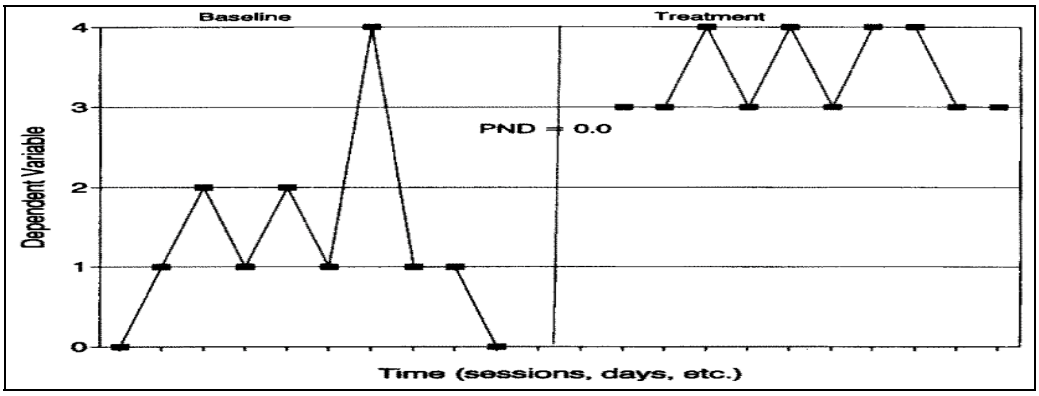

Fig 1: Data set that indicates an erroneous low treatment effect using PND.

From: Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, *31*(6), 621-631[19].

The authors also noted that PND is problematic when a treatment has a contra therapeutic effect, yet the presence of an outlier indicates that the treatment has a small, positive

effect. See Figure 2 for an example provided by the authors in which an intervention has a detrimental effect, yet the outlier indicates a small positive, therapeutic effect.

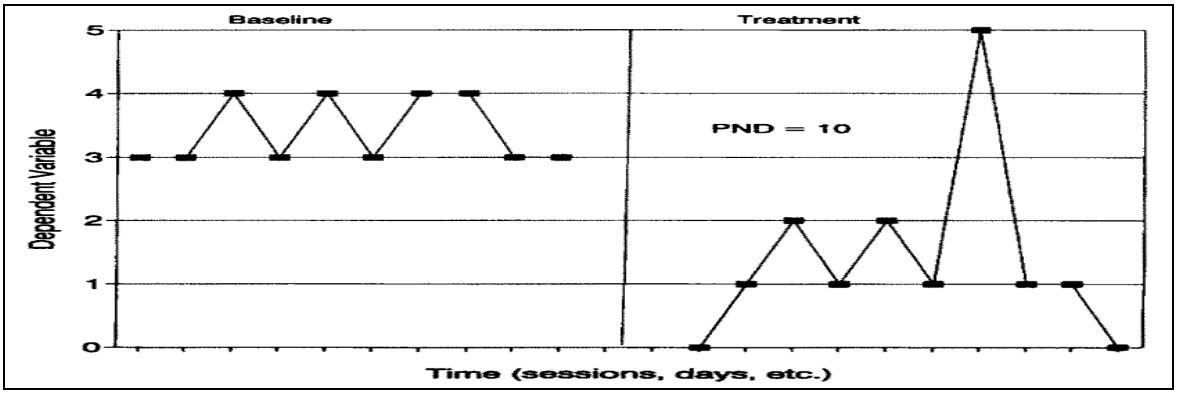

Fig 2: Data set that indicates an erroneous positive treatment effect using PND.

From: Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, 31(6), 621-631[19].

The authors surmised that PND is problematic when an existing trend is in effect during baseline and continues into the treatment condition (i.e., an extraneous variable is having an effect on the data). See Figure 3 for the authors' example graph that would yield a PND score of 100%, yet visual analysis would determine that the treatment alone was not responsible for the increased performance.

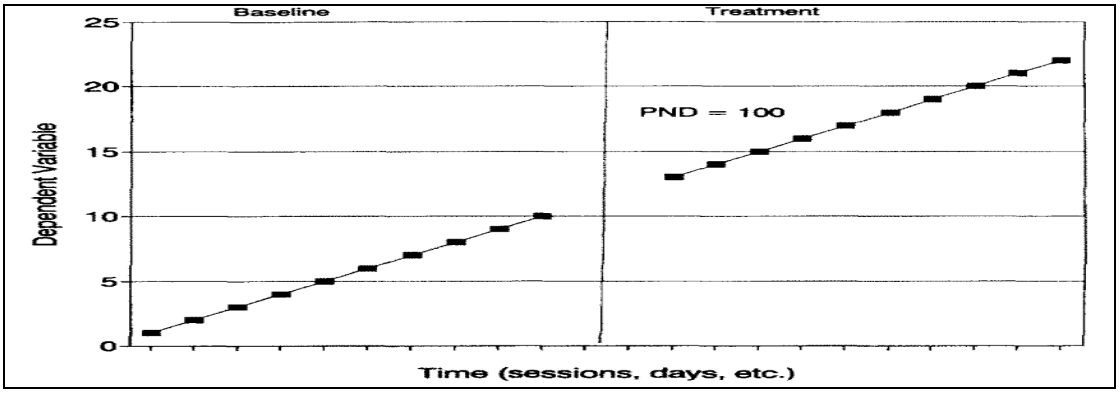

Fig 3: Data set that indicates an erroneous strong treatment effect using PND.

From: Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, *31*(6), 621-631[19].

The authors reported that PND might erroneously indicate no magnitude of effect when baseline data indicate a downward trend, but the intervention data indicate an upward trend. See Figure 4 for the authors' example of a graph that indicates 0% PND, yet clearly indicates a strong treatment effect using visual analysis.

6. Conclusions

The inclusion of effect size estimates has increasingly become a predominant component of research summaries for group design studies (Nakagawa & Cuthill, 2007) [1]; however, the inclusion of effect sizes in SCD is a contended issue (Campbell, 2004; Haardörfer, 2010) [15, 16]. A growing body of evidence indicates the potentially limited utility of PND as a metric for evaluating ES for SCD (Allison & Gorman, 1993; Campbell, 2004; Wolery *et al.*, 2010) [19, 15, 17]. Based on the literature reviewed herein, alternative metrics to PND should be considered to evaluate the magnitude of treatment effectiveness such as modified regression approaches or standard visual analysis (Allison & Gorman, 1993)[19]. The use of PND remains a controversial method of quantifying treatment effectiveness; therefore, researchers/practitioners should exercise caution when employing PND to evaluate treatment effectiveness. The possibility of using effect sizes in conjunction with visual analysis to augment research findings is potentially informative and valuable to SCD. Researchers should continue to investigate methods of logically applying statistical measures to SCD in a meaningful manner.

7. References

1. Kelley K, Preacher KJ. On effect size. Psychological methods, 2012; 17(2):137.
2. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews, 2007; 82(4):591-605.
3. Wilkinson L. American Psychological Association Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist. 1999: 54; 594-604.

4. Tuckman BW, Harper BE. Conducting educational research. Plymouth, MD: Rowman & Littlefield Publishers, 2012.
5. Huck SW, Reading statistics and research. Boston: Pearson, 2012.
6. Shaver JP. What statistical significance testing is, and what it is not. The Journal of Experimental Education. 1993; 61(4):293-316.
7. Gast DL, Spriggs AD. Visual analysis of graphic data. In D. Gast & J. Ledford (Eds.), Single Case Research Methodology: Applications in Special Education and Behavioral Sciences, 2e: Applications in Special Education and Behavioral Sciences New York, NY: Routledge. 2014, 176-210.
8. Parsonson BS, Baer DM. The analysis and presentation of graphic data. Single-subject research: Strategies for evaluating change, 1978, 105-165.
9. Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, *et al*. Single-case designs technical documentation. What Works Clearinghouse, 2010.
10. Horner RH, Carr EG, Halle J, McGee G, Odom S, Wolery M. The use of single-subject research to identify evidence-based practice in special education. Exceptional Children, 2005; 71(2):165-179.
11. Wolery M, Gast DL, Ledford JR. Comparison designs. Single case research methodology: Applications in special education and behavioral sciences, 2014, 297-345.
12. Gast DL, Ledford JR. Applied research in education and behavioral sciences. In D. Gast & J. Ledford (Eds.), Single Case Research Methodology: Applications in Special Education and Behavioral Sciences, 2e: Applications in Special Education and Behavioral Sciences. New York, NY: Routledge. 2014, 1-18.
13. Manolov R, Solanas A, Sierra V, Evans JJ. Choosing among techniques for quantifying single-case intervention effectiveness. Behavior therapy, 2011; 42(3):533-545.
14. Campbell JM. Commentary on PND at 25. Remedial and Special Education, 2013; 34(1):20-25.
15. Campbell JM. Statistical comparison of four effect sizes for single-subject designs. Behavior Modification, 2004; 28(2):234-246.
16. Haardörfer R. Concerns with using Cohen's D and PND in single-case data analysis. Focus on Autism and Other Developmental Disabilities, 2010; 25:125-127.
17. Wolery M, Busick M, Reichow B, Barton EE. Comparison of overlap methods for quantitatively synthesizing single-subject data. The Journal of Special Education, 2010; 44(1):18-28.
18. Scruggs TE, Mastropieri MA. Summarizing single-subject research issues and applications. Behavior Modification, 1998; 22(3):221-242.
19. Allison DB, Gorman BS. Calculating effect sizes for meta-analysis: The case of the single case∗. Behaviour Research and Therapy, 1993; 31(6):621-631.