Journal of School Psychology 52 (2014) 109–122

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Analysis and meta-analysis of single-case designs: An introduction☆

William R. Shadish

School of Social Sciences, Humanities and Arts, University of California, Merced, 5200 North Lake Rd, Merced CA 95343, United States

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form 28 November 2013
Accepted 30 November 2013
Available online 31 January 2014

Keywords:
Single-case designs
Statistics
Meta-analysis

ABSTRACT

The last 10 years have seen great progress in the analysis and meta-analysis of single-case designs (SCDs). This special issue includes five articles that provide an overview of current work on that topic, including standardized mean difference statistics, multilevel models, Bayesian statistics, and generalized additive models. Each article analyzes a common example across articles and presents syntax or macros for how to do them. These articles are followed by commentaries from single-case design researchers and journal editors. This introduction briefly describes each article and then discusses several issues that must be addressed before we can know what analyses will eventually be best to use in SCD research. These issues include modeling trend, modeling error covariances, computing standardized effect size estimates, assessing statistical power, incorporating more accurate models of outcome distributions, exploring whether Bayesian statistics can improve estimation given the small samples common in SCDs, and the need for annotated syntax and graphical user interfaces that make complex statistics accessible to SCD researchers. The article then discusses reasons why SCD researchers are likely to incorporate statistical analyses into their research more often in the future, including changing expectations and contingencies regarding SCD research from outside SCD communities, changes and diversity within SCD communities, corrections of erroneous beliefs about the relationship between SCD research and statistics, and demonstrations of how statistics can help SCD researchers better meet their goals.

1. Introduction

Single-case designs (SCDs) are widely used in a number of fields to assess the effects of interventions (Gabler, Duan, Vohra, & Kravitz, 2011; Shadish & Sullivan, 2011). They are used when the problem of interest has a very low base rate so that large numbers of units are difficult to locate, when the nature of the treatment requires a high degree of tailoring of treatment to the individual case, and when pilot work would be useful to demonstrate proof of concept prior to fielding a larger experiment. However, evidence from SCDs has not been widely used in reviews about evidence-based practice. A key reason for that is the lack of widely accepted and formally-developed statistical methods for the analysis and meta-analysis of such designs. The last decade has seen exciting progress towards remedying that problem. The five articles in this special issue of the *Journal of School Psychology* present a comprehensive sample of this work.

A key purpose of the special issue is to present these developments to the SCD research community in a manner that makes it possible for those researchers to learn them and try to use them in their work. So although the articles do present the statistical background and equations that represent their approaches, they also give extensive details about the computer programs and

☆ This research was supported in part by grants R305D100046 and R305D100033 from the Institute for Educational Sciences, U.S. Department of Education, and by a grant from the University of California Office of the President to the University of California Educational Evaluation Consortium. The opinions expressed are those of the author and do not represent views of the University of California, the Institute for Educational Sciences, or the U.S. Department of Education.

E-mail address: wshadish@ucmerced.edu.

ACTION EDITOR: Randy Floyd.

0022-4405/$ – see front matter
http://dx.doi.org/10.1016/j.jsp.2013.11.009

syntax that they use in doing the analyses. Some of these programs are familiar to most SCD researchers, such as SPSS and SAS, and others may require less commonly used software, such as R (R Development Core Team, 2012) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). SCD studies have usually not used statistics except for means or proportions, but for reasons discussed in this article, they may begin to use a wider array of statistics more often.

In each article, the analytic methods differ in approaches and assumptions, sometimes substantially. Yet all produce an effect estimate, sometimes standardized and sometimes not. Hence, the question arises whether these approaches all yield a similar answer. To help answer the question, all five articles apply their statistics to the same SCD study, a set of nine single-case ABAB designs from Lambert, Cartledge, Heward, and Lo (2006) on the effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students (Fig. 1). I digitized data for the nine cases using reliable and valid methods described elsewhere (Shadish et al., 2009) and then distributed the data to all authors. The results are summarized in the description of each article in the next section.

In addition, a few of the articles synthesize results over studies. Again, they use a common dataset, a group of six studies of the effects of Pivotal Response Training (PRT) on children with autism (Koegel, Camarata, Valdez-Menchaca, & Koegel, 1998; Koegel, Symon, & Koegel, 2002; Laski, Charlop, & Schreibman, 1988; Schreibman, Stahmer, Barlett, & Dufek, 2009; Sherer & Schreibman, 2005; Thorp, Stahmer, & Schreibman, 1995) and one study using the same methods and outcomes on adults (LeBlanc, Geiger, Sautter, & Sidener, 2007). To facilitate interpretation of some graphs in this special issue, the study identification number (SID) follows each of these references in the bibliography. Again, I digitized data from the articles so that all authors were analyzing the same dataset. To keep the dataset simple, it contains only outcomes related to child verbalizations (a bibliography showing which outcomes were kept is available from the guest editor), it does not include any maintenance/generalization/follow-up phases, and it only includes studies with at least three cases given that is the minimum number of cases needed in one of the articles in this special issue (Shadish, Hedges, & Pustejovsky, 2014-this issue) and we wanted all authors to analyze exactly the same data set. It happens that all the PRT studies used a multiple baseline design across cases, except Schreibman et al. (2009) that used a multiple baseline ABC design from which we deleted phase C in order to increase comparability to the other multiple baseline studies. In addition, the dataset contained three covariates that could be used as moderator variables: (1) Sex of participants (0 = male, 1 = female, 2 = both), (2) Age of child in years (using an average age if only that was reported), and (3) Location where the research was conducted (0 = Santa Barbara, 1 = San Diego, 9 = Other).

2. Brief introduction to the articles

The first article by Shadish, Hedges, and Pustejovsky presents a newly developed standardized mean difference statistic (d) for single-case designs that is in the same metric as the typical standardized mean difference statistic used in between-groups designs. It assumes normally distributed data and stationarity (no trend), and is corrected for small sample bias in the manner that is common in between-groups research, yielding Hedges' g. The authors have SPSS macros, and also graphical user (point-and-click) interfaces for the macros. The authors show how to compute the effect size for the Lambert et al. (2006) study, yielding standardized mean difference statistic of $g = 2.514$ ($s^2 = .0405$; 95% confidence interval $2.120 \leq \delta \leq 2.909$). So the number of intervals with a disruptive behaviors decreased by about two and a half standard deviations. The standard deviation is 2.16, so the decrease was about $2.16 \times 2.514 = 5.43$ fewer intervals with a disruptive behavior, generally consistent with visual analysis of the Fig. 1. Then, they show how to compute power analyses for this effect size using simple SPSS macros, which facilitates the planning of studies to have sufficient sensitivity to detect effects. Finally, they show how to conduct a meta-analysis on the PRT data set, finding that the random effects average effect size is $\overline{g} = 1.01, SE = .14, p < .001$. That is, PRT treatment produced an effect of about one standard deviation on the outcome measures, on average. They also demonstrate a wide range of meta-analytic techniques including influence diagnostics, forest plots, fixed and random effects meta-analyses, cumulative meta-analyses, moderator analyses, and publication bias analyses.

The second article is by Shadish, Zuur, and Sullivan. It concerns the key issue of linear and nonlinear trends in single-case design data. Many of the current effect size methods either ignore trend or explicitly assume no trend. Many general linear model approaches like regression and multilevel modeling can model trend but require the researcher to know the form of the trend—linear or nonlinear, and if the latter, how nonlinear. Unfortunately, the researcher rarely if ever knows the form; and if it is specified incorrectly, point estimates and standard errors will likely be wrong. To address this problem, Shadish, Zuur, and Sullivan introduce a semi-parametric method called generalized additive models (GAMs), which allows the data to prescribe the presence and shape of trends in the model. The authors show how to test a wide variety of models with GAMs. They do not assume normality because the outcome is a rate, so they use a binomial logistic model, adjusted for overdispersion. The logit effect size in their best fitting model on the Lambert et al. (2006) data was -3.347 ($s^2 = .830$). Interpreting this, they found a drop of about 6.7 intervals in which a disruptive behavior was observed. Then, these authors extend the analysis to generalized additive mixed models (GAMMs). These are the GAM equivalent of multilevel models that can allow modeling of autoregressive terms and random effects, although it is not clear whether single-case design data sets are large enough to support the computations. The authors conclude that GA(M)Ms can be used either as a primary analytic approach for SCDs, or as a means to examine the extent to which nonlinearities in data from SCDs might affect overall conclusions about treatment effectiveness. These authors did not do any meta-analytic work because they were not yet satisfied that an accurate effect size measure for non-normally distributed outcomes is available; but they discuss how such work could be done.

The third article by Rindskopf shows a fully Bayesian analysis of SCDs using WinBUGS (Lunn et al., 2000; Spiegelhalter, Thomas, Best, & Lunn, 2004), a common, free program for doing Bayesian statistics. Rindskopf begins with a discussion of the