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 © 2013 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
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 $\bar{g} = 1.01, SE = .14, p < .001$. 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

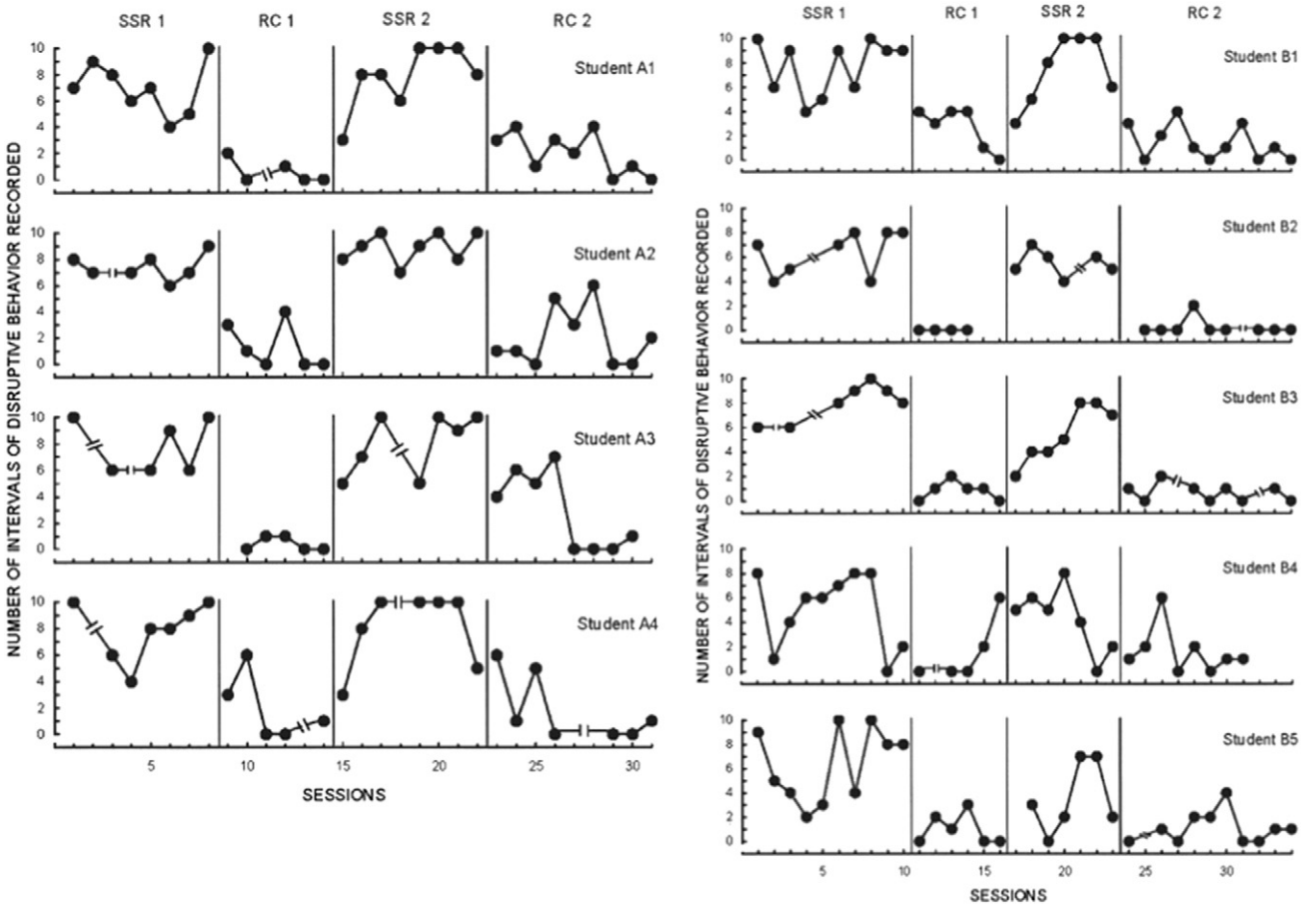

Fig. 1. Number of Intervals of Disruptive Behavior Recorded during single-student responding (SSR) and response card treatment (RC) conditions. Adapted from "Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students," by Lambert et al., 2006, *Journal of Positive Behavior Interventions*, 8e, pp. 94–95. Copyright 2006 by Sage Publications.

relative merits of multilevel modeling in the frequentist mode in which it is usually done (e.g., Shadish, Kyse, & Rindskopf, 2013) versus the Bayesian mode he presents. He argues that the Bayesian approach is particularly useful because it can cope with small samples better. He then shows how to model the Lambert et al. (2006) data in WinBUGS both without and with predictors, also assuming a binomial distribution for the outcome. He found a drop in the log-odds of disruptive behavior from phase A to phase B of -2.78, which translated to a drop from a .69 probability of disruptive behavior in phase A to a .13 probability of disruptive behavior in phase B—or about 5.7 fewer intervals of disruptive behavior. He also presents a second example using a dataset by Horner et al. (2005). That dataset requires more complex nonlinear trend specifications and accommodation of a floor and ceiling effect, which Rindskopf shows how to specify and program. This article did not do any meta-analytic work, although Rindskopf, Shadish, and Hedges (2012) have shown elsewhere how a standardized mean difference statistic could be computed using the methods Rindskopf illustrates in the present article.

The fourth article by Moeyaert, Ferron, Beretvas, and Van Den Noortgate shows how to apply multilevel analysis computed in SAS Proc Mixed (Little, Milliken, Stroup, Wolfinger, & Schabenberger, 2007). In the first part of the article, they use a two-level multilevel analysis to summarize results over multiple cases in one study using the Lambert et al. (2006) data set. They show how to examine a series of multilevel modeling choices, for example, whether or not to estimate autocorrelation, whether or not to include trends, whether to constrain the effect estimate for the second AB pair to be the same as that for the first AB pair, and alternative distributional assumptions. In one model using a binomial distribution, for example, they find a significant reduction in the number of intervals of challenging behaviors after treatment (e.g., treatment effect) of 5.61 (after translating back from logits to the original count of intervals); their other models yielded treatment effect estimates of about the same magnitude. In the second part of the article, they show how to use a three-level multilevel analysis to summarize results over multiple studies that each contains multiple cases, all measured on the same outcome measure—percentage of spontaneous or appropriate behavior in five studies from the PRT dataset but assuming those data are normally distributed. Across six different models, they found a fairly consistent and significant treatment effect of about 25% to 32% improvement in the outcome. Because they analyzed only a subset of the data analyzed by Shadish, Hedges, et al. (2014-this issue), we cannot compare the results from the two approaches.

The fifth article by Swaminathan, Rogers, and Horner describes a Bayesian generalized least squares approach to the analysis of SCD designs, giving careful attention to parameter interpretation, and approaches to estimating the autocorrelation. Then the authors propose an effect size measure that takes into account changes in slopes and intercepts in the presence of serial dependence and provides an integrated procedure for the analysis of SCDs through estimation and inference based directly on the effect size measure. They then present a multilevel model that is appropriate when several subjects are available, integrating this model into the Bayesian procedure to provide a standardized effect size measure comparable to the effect size measures in a between-subject design. Finally, they present an illustrative application of the Bayesian analysis using data from the Lambert et al. (2006) study assuming the outcome is normally distributed. The overall standardized effect size is −2.49 for the A1B1 comparison, 1.69 for the B1A2 comparison, and −2.33 for the A2B2 comparison.

So did the different analyses in these five articles produce the same answer when applied to the Lambert et al. (2006) data? The estimates were:

- Shadish, Hedges, and Pustejovsky: $g = 2.514$, or 5.43 fewer intervals of disruptive behavior.
- Shadish, Zuur, and Sullivan: 6.7 fewer intervals of disruptive behavior.
- Rindskopf: 5.7 fewer intervals of disruptive behavior.
- Moeyaert et al.: 5.61 fewer intervals of disruptive behavior
- Swaminathan et al: $g = 2.49$ for A1B1 and 2.33 for A2B2, or 5.38 and 5.03 fewer intervals of disruptive behavior, respectively.

The decrease in disruptive behavior ranged from 5.03 to 5.70 over four of these five analyses, not identical but reasonably consistent with each other and with visual inspection of the graph. The one modestly outlying effect, the 6.70 in Shadish, Zuur, and Sullivan, comes from the only analysis to model nonlinearities thoroughly, suggesting the possibility that such modeling may have a noticeable effect on the results.

These five articles are followed by three commentaries from SCD researchers (Maggin & Odom, 2014-this issue; Kratochwill & Levin, 2014-this issue; Fisher & Lerman, 2014-this issue). These authors were chosen to represent different parts of the SCD communities, ranging from researchers to reviewers to editors, some with quantitative interests and some without, but all widely experienced and deeply involved with SCDs.

3. Issues to keep in mind when reading the articles

It may help readers to read, conceptualize, and evaluate the five articles if they keep in mind the following issues. These are issues that must be dealt with by any analytic method for SCDs. In doing so, the intention is not to criticize any particular article or approach. After all, these analyses are developing rapidly to try to cope with the issues raised below, doing so takes time, and any single article cannot reasonably be expected to address all of them at the field's current state of development. Rather, the purpose is to educate and to point the way to the most important current and needed future developments.

3.1. Modeling trend

Trend refers to systematic changes in outcome observations over time within a case. When no trend exists, observations may fluctuate due to chance, but they do so around a horizontal line in both treatment and baseline conditions. Trend occurs when

observations increase or decrease systematically, either linearly or with nonlinearity that could take on a very wide range of functional forms—for example, quadratic, cubic, logistic, and so forth. Moreover, trend can be different in the baseline and treatment conditions.

Some methods such as Shadish, Hedges, et al. (2014-this issue) *d*-statistic assume that no trend exists, or alternatively, that the researcher has removed any trend from the data by one of several means before analyzing the resulting residuals. This assumption of no trend is true of nearly all of the effect size estimates in the literature (Parker, Vannest, & Davis, 2011). Assuming no trend is less desirable than modeling trend explicitly, although we have little clear empirical research on the prevalence and form of trend in the SCD literature.

The other four articles use methods that allow for trend, and fall into two categories. The parametric methods, exemplified by Rindskopf (2014-this issue), Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014-this issue) and Swaminathan, Rogers, and Horner (2014-this issue), require the researcher to specify the functional form of trend. Problematically, the functional form is in principle unknown, and misspecification of the functional form can result in erroneous effect estimates and standard errors. The semi-parametric methods (Shadish, Zuur, & Sullivan, 2014-this issue) combine parametric and nonparametric predictors. The nonparametric predictors allow the data to help specify the functional form by applying a smoothing function to the terms for time and for the interaction between time and treatment, but the researcher can still add parametric predictors. Simulation data (Sullivan, Shadish, & Steiner, in press) show that if the researcher actually does know the functional form, the parametric methods are more powerful and accurate. Otherwise, the semi-parametric methods generally perform better. Because the researcher rarely does know the functional form, this finding has potentially important implications for modeling trend more carefully in SCDs.

3.2. Modeling error structures

SCD researchers widely think of this issue as the problem of modeling the autocorrelation, that is, the serial dependencies of observations over time within cases. All the articles except Rindskopf take the autocorrelation into account in the error covariance structure. However, the problem is actually a bit more complex than simply modeling the autocorrelation. The reason is that multilevel models that include random effects for the intercept and slope, as Rindskopf does, also create an error covariance structure. After all, a benefit of multilevel modeling is said to be its ability to account for within case correlations. If so, one might ask what the point is of adding an autocorrelation estimate to a multilevel model with these random effects. So should the SCD researcher model random effects, or autocorrelation, or both? Two issues arise in answering this question: an issue of interpretation and an issue of statistics.

The two approaches—modeling autocorrelation versus random effects—imply different interpretations (Shadish, Kyse, & Rindskopf, 2013). Attributing error covariances to autocorrelation implies that the ordering of observations within phases is important so that observations cannot be randomly exchanged within phases. It also implies that a time dependent variable creates the error covariances, for example, a tendency to behave more similarly in sessions that are closer in time than in those further apart. Conversely, attributing error covariances to random effects implies that observations within phases are randomly interchangeable (e.g., one can replace the first treatment observation with any other randomly selected treatment observation), so that the order of observations within phases does not matter. It also implies that some error covariance is due to systematic differences among case characteristics, and those differences cause correlated errors.

It seems likely that both these explanations are true in SCD data—that is, some portion of error covariance is due to case characteristics, and some portion is due to time dependent processes. If so, one approach would be first to model random effects in order to account for those parts of the error covariances that are due to systematic differences between cases, and then add an autoregressive component to model the remaining time dependent error covariances. An interesting question is how one would empirically investigate both of these explanations. The answer would require measuring (a) variables at the case level that might predict random effects and (b) variables associated with time that might predict time-dependent processes.

The second issue is a practical one having to do with the resulting statistics. A key reason to model error structures is to ensure that the standard errors associated with effect estimates (and other predictors in the model) are reasonably accurate so that Type I error rates are well controlled. Preliminary evidence suggests that one can successfully do so either by modeling the random effects or by modeling the autocorrelation (Gurka, Edwards, & Muller, 2011; Hedeker & Gibbons, 2006; Shadish, Kyse, & Rindskopf, 2013; Singer & Willett, 2003). While effects of modeling errors in different ways need further study, the implication is that it may not be necessary to model the autocorrelation in models that include random effects and that the latter may control Type I errors reasonably well.

3.3. Standardized effect size statistics

Some models in this special issue result in a standardized effect size statistic. Although standardized effect sizes are not necessary to integrate cases measured on the same outcome within a study, they are crucial to integrating results over different outcome measures, whether within studies or across studies via a meta-analysis. The best developed effect size from a statistical point of view is the Shadish, Hedges, et al. (2014-this issue) *d*-statistic, which Hedges, Pustejovsky, and Shadish (2012, 2013) have shown is the same as the *d*-statistic typically used in the between groups literature and which appropriately takes into account the number of cases, the number of time points within cases, the ratio of between case variance to total (between plus within) variance, the lag-one autocorrelation, and in the case of reversal designs, the number of AB pairs for each case. The flaw in

it, as noted previously, is its failure to model trend and to allow for any distributions other than normal. The authors hope to remedy that flaw over the next few years, but the careful statistical development of such an effect size index takes time.

The remaining articles that present standardized effect size measures make significant progress towards remedying the problems of assuming no trend and a normally distributed outcome distributions. However, they do so with less formal grounding in statistical theory. They do not demonstrate, for example, that the resulting statistic is formally identical to the *d*-statistic from between groups studies, and the bias and efficiency of their estimates are not entirely clear. Further, although they may use more appropriate Poisson or binomial distributions to analyze data that use counts or proportions as outcomes, it is less clear that standardizing the resulting coefficients by between case variance results in a *d*-statistic because the latter inherently requires a normal distribution. No matter what approach one takes, however, the results are somewhat more likely to be valid when compared to past effect size statistics (Parker, Vannest, & Davis, 2011). The latter neither are formally grounded in statistical theory nor do they take trend and error covariances into account.

3.4. Outcome distribution assumptions

Until very recently, nearly all approaches to the analysis of SCDs have either ignored the outcome metric or assumed it is appropriately characterized by a normal distribution. Both are problematic because most outcomes in SCDs are either counts or rates (Shadish & Sullivan, 2011) that are best modeled by Poisson or binomial distributions, or extensions thereof (Shadish, Kyse, & Rindskopf, 2013).

Shadish, Hedges, et al. (2014-this issue) and Swaminathan et al. (2014-this issue) both assume normality. Shadish, Hedges and Pustejovsky assumed normality because normality made it more tractable to develop an effect size that is well-grounded in statistical theory; in principle, it should be possible to extend this work to other distributions. Swaminathan et al. focused on solving a different problem—the autocorrelation and trend—even though their methods could be extended to incorporate other outcome distributions. The remaining three articles do use other outcome distributions which are crucial because SCD researchers can make serious errors about the presence of an effect when they erroneously assume normality. For instance, when the outcome is measured by counting a behavior during a session, it may be best modeled with a Poisson distribution (or some variant of that like a negative binomial distribution). When outcome is measured as a proportion of trials in which the outcome was observed, a binomial (or related) distribution may apply. Shadish, Kyse, and Rindskopf (2013) elaborate some of the other distributional assumptions and when they might apply to SCD research. Different distributions, in turn, have implications that are not usually encountered with normal distributions. In a Poisson distribution, for instance, the mean and the variance are equal. So if, for example, an increasing linear trend results in a higher mean count in a treatment phase than in baseline in a multiple baseline design, the variation around that count will also be larger, with more extreme observations potentially both above and below the mean. This equality of mean and variance may lead the researcher to incorrectly conclude an effect is present, or not, based on thinking that such extreme observations reflect a treatment effect when they really reflect chance. Such errors particularly affect the visual analysis of SCDs, so the recent movement towards more appropriate distributional assumptions in SCDs is vital.

3.5. Bayesian estimation

Both Rindskopf (2014-this issue) and Swaminathan et al. (2014-this issue) use Bayesian estimation as part of their analyses. Bayesian statistics are likely to be useful in SCD research for two reasons. The first and most important reason is that Bayesian statistics work very well with small sample sizes. Small sample sizes characterize SCDs in general as most studies contain relatively few cases, and many cases contain relatively few observations over time. Similarly, estimating the autocorrelation is biased in small samples, and it may be that Bayesian estimates will prove more accurate than other estimates. At a minimum, Bayesian estimates will reduce the role that sampling error plays in the magnitude of the autocorrelation (Shadish, Rindskopf, Hedges, & Sullivan, 2012). The second reason is that Bayesian statistics take into account the researcher's prior knowledge about the uncertainty about parameters such as the autocorrelation, the intraclass correlation, the effect size, random effects, and errors. Although the past reliance on visual analysis in SCD research means we have little such prior knowledge, that knowledge would not be difficult to accumulate as statistics become more widely used. The result should be more accurate and powerful estimates.

3.6. Power analyses

Only one of the articles (Shadish, Hedges, et al., 2014-this issue) presents a way of doing statistical power analyses for SCDs. Yet power analyses will prove particularly useful to SCD research in two related situations. The first is in planning studies with enough cases and time points to ensure an effect size of interest can be detected if it exists. The second is in submitting grant applications to funding agencies. All other things being equal, such agencies, and the researchers who review for them, seem likely to give preference to grants that add a power analysis compared to those that do not. For the articles that present multilevel models but not power analyses for them, it may be that existing software such as Optimal Design (Spybrook et al., 2011) can be adapted to the problem.

3.7. Accessibility

All of the five articles use complex statistics. SCD researchers are unlikely to use them unless they are made accessible through the use of clearly annotated syntax, macros, or graphical user interfaces (GUIs). Only Shadish, Hedges, et al. (2014-this issue) have macros and GUIs, both for computation of the *d*-statistic and for estimating power; they also have an accompanying manual. The remaining four articles present annotated syntax, but GUIs would help even more. It may or may not prove feasible to create such interfaces for programs like WinBUGS, but we cannot know until we try. When these and other authors submit grant proposals to continue their work, therefore, it could prove very beneficial if the budgets include funding for a programmer who can do this work. GUIs have the added advantage that they help prevent users from experiencing frustrating syntax errors. The reason is that GUIs generally create syntax in the background, and they are programmed to do so correctly. So they bypass a step in the process that can lead to frustration and errors among users.

4. Quantitative analyses and SCD tradition

The primary intent of this special issue is to describe and demonstrate cutting-edge approaches to quantitative analysis in SCDs. Yet most applied research studies using SCDs do not include much quantitative analysis beyond simple descriptive statistics like means or proportions. One reason has to do with the early history of SCD research. Some of the most important early writers (e.g., Sidman, 1960; Skinner, 1938, 1972) argued that Fisherian group experiments and their associated statistics were inimical to intimate interplay between experimenter and subject that characterizes SCD research and that they can hinder the development of the good experimental judgment by fostering undue reliance on simple statistical decision rules that do not always accurately reflect the SCD process.

Following this early lead, many SCD researchers object to quantitative analysis as the imposition of "group-statistical" methods (Perone, 1999, p. 111) on SCD researchers who do not do group research or who do not wait until a study ends to look at the data. They prefer visual methods to statistical methods based on the rationale that visual analysis is less likely to say a treatment works when it does not, even at the risk of overlooking some potentially effective treatments. Of course, SCD researchers phrase this rationale in terms of functional relationships rather than treatment effects. Baer (1977), for example, says visual analysis is unlikely "to affirm that a certain variable is a functional one, when in fact it is not" (p. 170), and is more likely "to deny that a certain variable is a functional one, when in fact it is" (p. 170). SCD researchers are sometimes also skeptical that statistical analyses can, in the words of one reviewer of a manuscript of mine submitted to this special issue, capture fully all the aspects of visual analysis needed to infer experimental control, such as level, trend, variability, overlap, immediacy of effect, and phase consistency.

The arguments mounted by all these authors and others (e.g., Parker & Vannest, 2012; Salzberg, Strain, & Baer, 1987) should be required reading for all of us who propose statistical methods for SCD research. They remind us not only that we might not always be welcomed with open arms but also more importantly that we need to learn about the historical bases and deep philosophical rationales for principled and thoughtful doubt about what statistical analysis can contribute to SCD research. Statistical methods are more likely to be perceived as useful in the SCD community when they respect these concerns, respond to SCD researcher needs, build on those needs where possible, and demonstrate the value added by new statistics that sometimes lack immediately obvious connection to the SCD process. That being said, four factors give cause to think that quantitative methods may have an increasing role in SCD research: changes and diversity within SCD communities, changing expectations and contingencies regarding SCD research from outside 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

4.1. Changes within the SCD community

The first factor concerns changes that are occurring, however slowly, in the SCD community itself. Some anecdotal evidence suggests that some SCD researchers see statistical analysis as a way to improve the field while at the same time making their mark on it. At the same time, however, they are restrained somewhat by the worry about how statistical analysis will be received by some SCD leaders who remain opposed to employing them. Researchers open to statistics include both junior and senior members of the community. To the extent that they include more junior members, history of science suggests that generational turnover is a common way that fields change.

In addition, using the term "SCD community" is really a misnomer. Rather, we should speak of SCD communities in the plural. Some parts of the many SCD communities are probably more opposed to statistical analysis than other parts. After all, researchers who publish in, say, the *Journal of Applied Behavior Analysis* may have very different views on the matter than researchers who publish in, say, *Neuropsychological Rehabilitation*.

The last several decades have also seen modestly increased use in SCD communities of the overlap statistics (Parker, Vannest, & Davis, 2011; Scruggs & Mastropieri, 2013) for meta-analytic purposes. A review of the 2008 database compiled by Shadish and Sullivan (2011) found these statistics are even sometimes used in the analysis of SCD primary studies (Burns, Peters, & Noell, 2008; Fudge et al., 2008; Ganz & Flores, 2008; Ganz, Kaylor, Bourgeois, & Hadden, 2008; Lane et al., 2008), along with the occasional use of some version of a *d*-statistic (Scattone, 2008). So the seeds of changes that might encourage the use of quantitative methods are already present within SCD research itself.

4.2. Changes in external contingencies

The second reason to think quantitative methods will be used more often is that external environmental contingencies may reinforce increased use of statistical methods. The three main contingencies concern the expectations of the evidence-based practice community, of granting agencies, and of journal editors. As an example of the first point, when the What Works Clearinghouse (WWC) Technical Advisory Group first considered which studies could count towards providing good evidence, SCDs were rejected largely for lack of statistical analyses. Indeed, it was that fact that encouraged me and some others to pursue such analyses starting 10 years ago because we wanted SCD research included. Today, WWC incorporates SCDs into a few of its reviews based on a visual analysis protocol (Hitchcock et al., in press; Kratochwill et al., 2013), but it remains very interested in statistical innovations that might be useful in evaluating SCD effects and their magnitudes. The more SCD researchers can speak to the evidence-based practice community using common statistical parlance, the more likely SCD research is to be used in such reviews.

The second set of external expectations comes from agencies that fund SCD research, like the Institute of Education Sciences. Such agencies also fund research on the statistical analysis of such designs. It seems reasonable that they may eventually expect those analyses to be used, and so to encourage SCD researchers to include planned data analyses in grant proposals. Ambitious grant applicants also seem likely to anticipate that eventuality by including a section on statistical analysis in order to try to gain a competitive edge over other applications. Even more, some of the articles in this special issue propose power analyses, and those same ambitious researchers may include them to help gain another edge.

Third, the editors of at least some of the journals that publish SCD research have expressed, more or less directly, a desire to see better statistical analyses of SCD data. Examples include the editors of *Journal of School Psychology* (Floyd, 2012, 2013, in press), *School Psychology Quarterly* (Kamphaus, in press), and *School Psychology Review* (Burns, in press). As anyone who has been an editor knows only too well, translating editorial opinions into research and publication practices is fraught with obstacles. Nonetheless, editors can influence those practices by choice of associate editors, editorial board members, and article reviewers. Very slowly over time, those choices can make a difference to what gets published.

4.3. Understanding erroneous arguments against statistics

A third factor that may contribute to more use of statistical analysis is that many of the reasons offered against statistical analysis are ill-founded. The implication is that SCD researchers and statistical analysts like those in this special issue may have far more in common than some previous writers have suggested. Consider some examples. Baer (1977) claimed that SCD research makes fewer Type I and more Type II errors than group researchers. However, no evidence exists that this claim is actually true. We can study it by comparing the narrative conclusions of SCD researchers to the statistical conclusions from analyses. In one case where I have done so (Shadish, Kyse, & Rindskopf, 2013), the two kinds of conclusions matched up very well. Also, the conclusions of Lambert et al. (2006) seemed to match up well with the statistical conclusions in the five articles in the special issue. One or two such examples are certainly not definitive, but they do suggest that claims need to be tested in data before being widely promulgated as fact.

Perone (1999) claimed that Skinner rejected group-statistics methods because he found that they insulated the researcher from the behavior of the subject. However, a close read of the pertinent quotation from Skinner shows he did not reject the statistic; rather, he rejected the design. Here is what Skinner said about his experience with group randomized experiments:

> You cannot easily make a change in the conditions of an experiment when twenty-four apparatuses have to be altered. Any gain in rigor is more than matched by a loss in flexibility. We were forced to confine ourselves to processes which could be studied with the baselines already developed in earlier work. We could not move on to the discovery of other processes or even to a more refined analysis of those we were working with. No matter how significant might be the relations we actually demonstrated, our statistical Leviathan had swum aground. (Skinner, 1956/1972, pp. 113–114)

Making changes in apparatuses, having flexibility, not being able to have extended baselines, and discovering new processes, are all matters of design, not statistics. Of course, it follows that having rejected the design, Skinner understandably had no use for the pertinent statistics. However, no modern analyst is suggesting SCD researchers should analyze single-case data with statistics appropriate for groups—that would make no statistical sense. Rather, the aim is to develop statistics that are appropriate for SCD research as it is done.

Criticizing Fisherian methods is not only beside the point, it forgets that one of the most famous experiments that Fisher analyzed is a single-case design: the lady tasting tea experiment (Salsburg, 2001). In Cambridge, England, during one summer in the 1920s, a group of men and women were having tea. One woman said tea with milk tasted different depending on whether the milk or the tea was poured first. Fisher was present, and after listening to the discussion among the group members, he arranged an experiment in which the lady was presented with eight cups of tea sequentially, in blinded and random order, four with milk poured first and four with tea poured first. He recorded her responses without comment. Apparently, she identified each one correctly (Salsburg, 2001) according to the report of a statistician who was present. Fisher later used this example to illustrate the use of randomization tests in his seminal book on experimental design (Fisher, 1935). Single-case designs are entirely compatible with at least some Fisherian statistics because not all Fisherian statistics are "group-statistical" (Perone, 1999, p. 111).

Perone (1999) also argued sensibly that "averaging data over many subjects can hide a multitude of sins" (p. 112). This argument is most assuredly correct. Yet, for two reasons, this argument does not imply that averaging data over many subjects

has nothing to offer at all. First, in the majority of the analyses described in this special issue, such averaging occurs only in addition to extensive analysis of individual cases. The choice is not one or the other when both can be done simultaneously. Second, averages are useful summaries even when not accompanied by individual analyses. In economics, the summary we call the gross national product (GNP) hides wide variability in the very diverse economic productivity and changing trends across sectors and firms of different sizes with different products, but no one thinks we should not use the summary because of that fact. More relevant to SCD research is the need in evidence-based practice to compare and summarize results over many different kinds of designs and across different kinds of treatments, cases, settings, and outcomes. Certainly, tension exists between thoughtful conceptual objections to averaging versus the desire to have SCD evidence admitted to the evidence-based practice oeuvre. Yet the world of evidence-based practice is one where averages and summaries are necessary because policymakers cannot prescribe interventions for each individual. Room exists for all these analyses.

Perone (1999) also argued that past criticisms of visual analysis (e.g., DeProspero & Cohen, 1979; Knapp, 1983; Matyas & Greenwood, 1990) often fail to take into account that statistical analysis is done after the end of a study but visual analysis is ongoing throughout the experiment. Many group researchers will not have previously understood this argument. Yet it is problematic on two counts when taken as an argument against statistical analysis. First, little doubt exists that visual analysis is, in fact, used at the end of SCD studies to aid conclusions about whether an intervention is effective. If so, it is fair to study the relative merits and flaws of visual and statistical analyses at that time for that purpose. Second, statistical analysis is also often ongoing throughout an experiment. An example is when independent review boards periodically examine data from randomized trials to decide whether to stop the experiment because it is causing harm or has such positive effects that it would be unethical to continue the study. Admittedly, ongoing visual analysis in SCDs plays a different kind of role than ongoing analysis in group studies, and is probably more thorough and frequent than in group experiments. Still, this example indicates how we all would benefit from continued dialog to help improve the precision and accuracy of arguments, and thereby perhaps arrive at better conclusions about the relative roles of visual and statistical analyses.

Salzberg et al. (1987) objected to summaries over cases within studies and over studies (i.e., meta-analysis) using overlap statistics on four counts. First, such statistics obscure patterns over time. True, but within-study summaries can be accompanied by within-study examinations of patterns over time, between-study summaries can sometimes also be done that way (Moeyaert et al., 2014-this issue; Swaminathan et al., 2014-this issue), and between-study summaries can still be useful even when they obscure patterns, as described previously.

Second, Salzberg et al. (1987) argued that summaries may miss vital idiosyncrasies within and between studies. They illustrated this claim by showing how simple summaries "failed to capture the relevant information or the issues" (p. 45) in six of the studies in a meta-analytic review. Indeed, higher-order (meta) summaries will inevitably have this effect because their very point is to abstract some key results without repeating every single result that a researcher could glean by reading the original articles. Yet this claim does not logically have any bearing on the fact that those summaries still contain relevant information that is not present in the within-case analyses.

Third, Salzberg et al. (1987) claimed that some syntheses they reviewed misrepresented facts or outcomes. Sadly, this observation seems to be an inevitable if undesirable aspect of science from time to time. Rarely is it intentional, but mostly it seems to be a function of the sorts of biases that are an all-too-inevitable part of being human, such as confirmation biases (Lilienfeld, Ammirati, & David, 2012). Science is a profoundly human endeavor, imperfect in many ways. That being said, it seems quite unlikely that SCD research is immune to such problems, or that by simply avoiding research synthesis SCD researchers will thereby avoid misrepresenting facts or outcomes themselves from time to time.

Fourth, Salzberg et al. (1987) claimed that such syntheses inappropriately draw conclusions about the relative merits of different interventions because studies of those interventions may differ in other important ways besides the intervention themselves. This point is an excellent one. Indeed, meta-analysts have known for decades that meta-analytic data are correlational data because, for example, studies are not randomly assigned to levels of all the covariates in a meta-analysis (e.g., Shadish, 1992). Thoughtful meta-analysis recognizes and studies such confounding variable to the extent possible, though it is never fully possible to know all the potential confounding covariates. It is best to treat most conclusions about the relative merits of different interventions as empirically grounded hypotheses for future research (but see Shadish, 1992, for some exceptions). Furthermore, this fourth criticism is not just true of quantitative syntheses. It is also true of any narrative review of a literature, and SCD researchers do such narrative reviews themselves (e.g., Datchuk & Kubina, 2013; Hagopian, Rooker, & Rolider, 2011; Martinez & Betz, 2013; McMillan, 2013; Payne & Dozier, 2013; Schertz, Reichow, Tan, Vaiouli, & Yildirim, 2012; Waldron, Casserly, & O'Sullivan, 2013).

A different view comes from Parker and Vannest (2012). These authors are certainly not against quantitative analysis of SCDs because they have been key pioneers in developing the nonoverlap effect sizes for SCDs (e.g., Parker, Hagan-Burke, & Vannest, 2007; Parker, Vannest, & Davis, 2011; Parker, Vannest, Davis, & Sauber, 2011). They make a very reasonable observation that "the statistical sophistication of omnibus models seems to be outpacing two-way communication with interventionists and behavior analysts, who are the only people with intimate understanding of the SCR design and data" (p. 256). We would like to hope that this special issue can be a vehicle for such dialog, and we suspect the quantitative analysts in this special issue would welcome opportunities for such communication. Concrete suggestions on how to make that happen would be helpful. Examples might be publishing practical examples in journals read by SCD researchers, holding workshops to familiarize SCD researchers with various quantitative methods (from which analysts can, in turn, learn more about SCD research), giving joint symposia at pertinent conferences, and holding special conferences bringing the two groups together to help make progress.

Parker and Vannest (2012) also note that many analyses of the kind represented in this special issue "are not easily understood and may require additional training to calculate or even to consume" (p. 255). No doubt this statement is true. Yet

that is no reason to avoid such analyses if they are the best model for the job. To paraphrase a well-known saying, "a statistical analysis should be made as simple as possible, but not simpler".[1] If we really desire an analytic method capable of modeling level, trend, variability, overlap, immediacy of effect, and phase consistency, it is unlikely an overly simple statistic will always do that job.

Further, it may be a mistake to underestimate the ability of SCD researchers to learn more complex statistics or apply the complex statistics they already know but rarely apply to SCDs. I have been pleased to hear from a number of them who are excited about the possibilities they see. Of course, a reciprocal obligation exists for quantitative analysts to make their analyses more accessible with examples, syntax, and graphical user interfaces, as we noted earlier. Still, when you combine this change within the SCD community with the changing external contingencies noted earlier, we are more optimistic that two-way communication (even collaboration!) will eventually result in potentially fundamental changes that may affect training programs and publication practices, as well as grant proposal submissions and reviews, and contributions to evidence-based practice.

Finally, Parker and Vannest (2012) repeat no less than five times the importance of keeping "the knowledgeable behavior analyst in control" (p. 264) of the data-analysis process. The problem with this hope is SCD researchers have already made their data public in published graphs. As is done in this special issue, nothing can stop scholars from digitizing those graphs and doing their own analyses of the resulting data. The problems are intellectually interesting, and funding agencies will give grants to develop and do analyses. Rather than trying to keep the behavior analyst in control, we should be trying to foster the kind of two-way communication that Parker and Vannest so rightly advocated.

4.4. Increased appreciation of what statistics can offer

Fourth, increased use of quantitative analyses in SCD research could also be fostered by clearer demonstrations of the benefits of using statistical methods. To introduce the first benefit, consider Fig. 2. The six graphs in that figure show simulated data. The outcome in each case is a count, simulated as overdispersed given how common that seems to be in SCD data (Shadish, Kyse, & Rindskopf, 2013; Shadish, Zuur, et al., 2014-this issue; Sullivan et al., in press). See Shadish, Zuur, et al. (2014-this issue) for detailed discussion of overdispersion. The simulation assumes an autocorrelation of .60. So that the interested or skeptical reader can replicate the general point about to be made, here is the code for generating and plotting the data in the free computer program R (R Development Core Team, 2012):

```
N <- 20
x <- c(1:20)
rho <- 0.6
log.lambda <- 1+arima.sim(model=list(ar=rho), n=N)
y <- rpois(N, lambda=exp(log.lambda))
df <- data.frame(x,y)
plot(x,y)
```

It is not hard to imagine a researcher looking at the data in Fig. 2 and concluding that a treatment is effective, either decreasing an undesirable behavior from baseline highs or increasing a desirable behavior from baseline lows. In fact, however, each of the graphs in Fig. 2 contains data points that are entirely due to chance (given the autocorrelation and the mean of the distribution). Not every simulation using the above code results in such graphs, but I found out about one out of five did. Try it! Of course, these graphs do not take into account that the SCD interventionist chooses the time point at which to intervene, which might not match well or at all the changes in trend and level seen in Fig. 2. Nor do the graphs take the design into account. The graphs might be fit into an AB or multiple baseline design, but it is much harder to generate chance data that mimics treatment effects in ABAB designs. So the chances of mistaking chance for a treatment effect in count data are much lower than one in five. Still, it should give pause to the thoughtful SCD researcher to realize exactly how much chance count data can resemble functional relationships. Arguably, then, the most important benefit of quantitative data analysis is the ability that statistics give the researcher to identify and model chance.

Second, statistics can provide information about various distributions appropriate for the kinds of outcomes used by SCD researchers. Most researchers know only the normal distribution, but count data may require a Poisson distribution, and rate data (number or percent correct out of a fixed number of trials) may require a binomial distribution. These distributions have quite different characteristics than normal distributions. For example, whereas a normal distribution has independent mean and variance, the variance of a Poisson distribution equals its mean. Also, Poisson and binomial distributions can be overdispersed, with more variance than the distribution theory predicts, requiring discussion of appropriate distributions for overdispersed data, such as negative binomial, beta-binomial, or zero-inflated distributions. Yes, one can hear the skeptic say that such discussions and analyses are too complex for SCD researchers. Yet it is precisely the failure to understand the characteristics of a Poisson distribution (variance equals mean) that can lead to mistaking chance for treatment effects in Fig. 2.

An implication of distribution theory is for overlap statistics. Some portion of the nonoverlapping data in Poisson distributed data is not caused by a treatment effect but by chance. After all, if treatment (or increasing trend in the absence of a treatment

[1] The original saying substitutes the word everything for the phrase a statistical analysis. The saying is often attributed to Albert Einstein, but does not appear in his work. See http://quoteinvestigator.com/2011/05/13/einstein-simple/ for details.

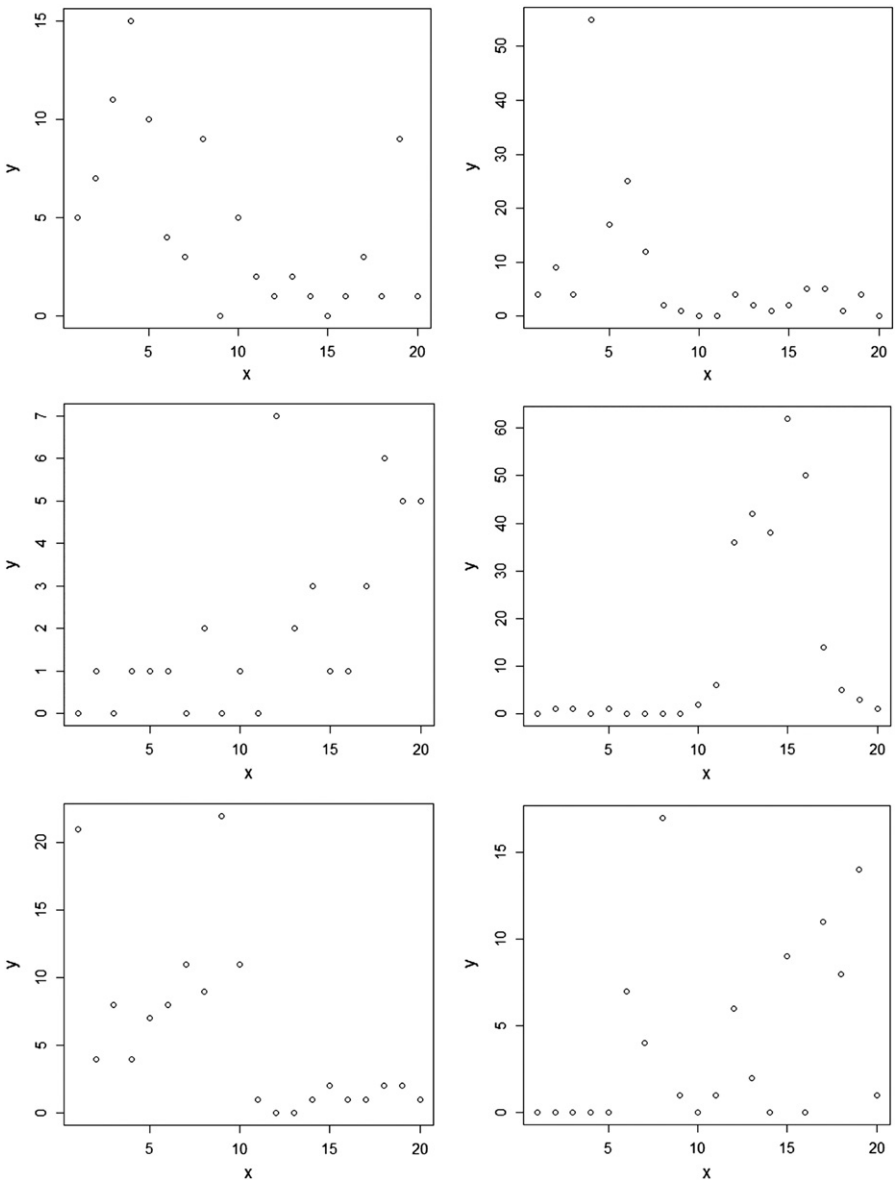

Fig. 2. Simulated SCD data using overdispersed Poisson distributions with an autocorrelation of .60.

effect) increases the mean during that phase, then it also increases the variance (Poisson variance equals mean), leading to more nonoverlapping data by chance. Statistics quantify this relationship between mean and variance in a more formal way, leading to formally developed standard errors appropriate for the kind of outcome data being gathered. That, in turn, leads to significance tests, confidence intervals, and the use of a host of statistics in both primary and meta-analysis that are now nearly completely absent in SCD publications but that have much to offer.

A third benefit is that statistics provide tools that allow SCD researchers to test some of the hypotheses that are part of the culture of SCD research. A good example is the previous suggestion to test Baer's (1977) claim that SCDs yield fewer Type I errors but more Type II errors. How could one test that claim without statistics? That is, how can we know that a functional relationship exists but the SCD researcher has incorrectly concluded it does not exist (a Type II error)? How would we know that an SCD researcher's conclusion that a functional relationship is present is, in fact, incorrect (a Type I error)? Using statistics to compare results from visual and statistical analyses is one useful way to do so.

A fourth benefit is that statistical analysis provides discipline to the process of drawing summative conclusions about the presence of a functional relationship at the end of a study. Perone (1999) is quite right to note that visual analysis is ongoing throughout a study, not just something done at the end. Nonetheless, when I reviewed more than one hundred SCD studies in our 2008 database (Shadish & Sullivan, 2011), every one also used visual analysis at the end to draw summative conclusions. Of these,

only one (van Oorsouw, Israel, von Heyn, & Duker, 2008) used any sort of visual analysis protocol at all to discipline the process. Given that we know factors like confirmation bias exist in scientists just as in all humans and that studies of visual analysis do suggest it has problems (e.g., DeProspero & Cohen, 1979; Knapp, 1983; Matyas & Greenwood, 1990), statistics have something to offer in disciplining a final summative judgment. To be clear, all quantitative analysts recognize that statistics are not perfect, and much more study is needed to find their strengths and weaknesses when applied to SCDs. So final judgment about functional relationships should never depend only on statistics. Even so, excluding statistics as evidence towards that judgment seems to make summative judgments more likely to be biased and not less likely.

Fifth, statistics can always in principle, and increasingly in practice, quantify the very qualities in data that SCD researchers use in deciding if a functional relationship is present—level, trend, variability, overlap, immediacy of effect, and phase consistency. Not all analyses quantify each and every one of these qualities, of course. An effect size rarely quantifies all of them, and the analyses that do quantify all of them may also be less immediately accessible and more difficult to learn, like the generalized additive models in Shadish, Zuur, et al. (2014-this issue). Yet such problems are remediable given time. The use of statistics in SCD research has been so rare for so long that it will take time to develop the optimal, accessible approach. We will have to learn to walk before we can run.

5. Conclusion

The most interesting question, to me at least, is not whether we will eventually develop statistical methods for SCDs that meet all of the desiderata mentioned in this article. I have no doubt we will do so—perhaps as soon as in the next five years. Rather, the most interesting question may be whether SCD researchers will use them. Despite all the reasons outlined above for thinking they might do so, doing so would also be a modest but real paradigm shift in SCD research. Such paradigm shifts are rare in science and nearly impossible to predict with accuracy. After all, in this case the shift depends on many factors that are not in any one person's or group's control, such as compelling contingencies and reinforcements from external sources, changes within the SCD communities in training and publication practices, recognition of SCD researcher misconceptions about statistics and of what statistics has to offer, generational turnover, and more.

Despite these obstacles, three reasons lead me to think that shift will occur quickly, if it has not begun to occur already—the analysis and meta-analysis of SCD research are intellectually interesting, are fundable, and SCD researchers have already placed their data in the public domain. So either SCD researchers will join in, or others will get the funding and publications. The latter could happen, resulting in one group that produces SCD research and one group that analyzes it. I bet that will not happen. I bet SCD researchers will see the advantages of collaborating with quantitative researchers to the benefit of all.

References

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, *10*, 167–172.

Burns, M. W. (2014). Reactions from the field: Single case studies in School Psychology Review. In T. K. Kratochwill, & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, D. C.: American Psychological Association (in press).

Burns, M. K., Peters, R., & Noell, G. H. (2008). Using performance feedback to enhance implementation fidelity of the problem-solving team process. *Journal of School Psychology*, *46*, 537–550.

Datchuk, S. M., & Kubina, R. M. (2013). A review of teaching sentence-level writing skills to students with writing difficulties and learning disabilities. *Remedial and Special Education*, *34*, 180–192. http://dx.doi.org/10.1177/0741932512448254.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis*, *12*, 573–579.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fisher, W. W., & Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: Lies, Damn Lies, and Statistics". *Journal of School Psychology*, *52*, 135–140 (this issue).

Floyd, R. G. (2012). A golden anniversary: Celebrating successes and establishing a vision for the future of the Journal of School Psychology. *Journal of School Psychology*, *50*, 1–6.

Floyd, R. G. (2013a). Enactment and evolution of the vision of the future of the Journal of School Psychology. *Journal of School Psychology*, *51*, 261–266.

Floyd, R. G. (2014). Reactions from the field: Single-case studies in the Journal of School Psychology. In T. K. Kratochwill, & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, D. C.: American Psychological Association (in press).

Fudge, D. L., Skinner, C. H., Williams, J. L., Cowden, D., Clark, J., & Bliss, S. L. (2008). Increasing on-task behavior in every student in a second-grade classroom during transitions: Validating the color wheel system. *Journal of School Psychology*, *46*, 575–592.

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care*, *49*, 761–768.

Ganz, J. B., & Flores, M. M. (2008). Effects of the use of visual strategies in play groups for children with autism spectrum disorders and their peers. *Journal of Autism and Developmental Disorders*, *38*, 926–940.

Ganz, J. B., Kaylor, M., Bourgeois, B., & Hadden, K. (2008). The impact of social scripts and visual cues on verbal communication in three children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *23*(2), 79–94. http://dx.doi.org/10.1177/1088357607311447.

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, *30*, 2696–2707. http://dx.doi.org/10.1002/sim.4293.

Hagopian, L. P., Rooker, G. W., & Rolider, N. U. (2011). Identifying empirically supported treatments for pica in individuals with intellectual disabilities. *Research in Developmental Disabilities*, *32*, 2114–2120.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* New York, NY: Wiley.

Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods*, *3*, 224–239.

Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*. http://dx.doi.org/10.1002/jrsm.1086.

Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2014). The what works clearinghouse single-case design pilot standards: Who will guard the guards? *Remedial and Special Education* (in press).

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179.

Kamphaus, R. W. (2014). Reactions from the field: Recommendations for single-case researchers. In T. R. Kratochwill, & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances*. Washington, D.C.: American Psychological Association (in press).
Knapp, T. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment*, *5*, 155–164.
Koegel, L. K., Camarata, S. M., Valdez-Menchaca, M., & Koegel, R. L. (1998). Setting generalization of question-asking by children with autism. *American Journal on Mental Retardation*, *102*(4), 346–357 (SID = 11).
Koegel, R. L., Symon, J. B., & Koegel, L. K. (2002). Parent education for families of children with autism living in geographically distant areas. *Journal of Positive Behavior Interventions*, *4*, 88–103 (SID = 8).
Kratochwill, T. R., & Levin, J. R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology*, *52*, 123–127 (this issue).
Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38. http://dx.doi.org/10.1177/0741932512452794.
Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8e*, 88–99.
Lane, K. L., Harris, K. R., Graham, S., Weisenbach, J. L., Brindle, M., & Morphy, P. (2008). The effects of self-regulated strategy development on the writing performance of second-grade students with behavioral and writing difficulties. *The Journal of Special Education*, *41*, 234–253. http://dx.doi.org/10.1177/0022466907310370.
Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, *21*, 391–400 (SID = 15).
LeBlanc, L. A., Geiger, K. B., Sautter, R. A., & Sidener, T. M. (2007). Using the natural language paradigm (NLP) to increase vocalizations of older adults with cognitive impairments. *Research in Developmental Disabilities*, *28*, 437–444 (SID = 16).
Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology*, *50*, 7–36.
Little, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2007). *SAS for mixed models.* Cary, NC: SAS Institute Inc.
Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.
Maggin, D. M., & Odom, S. (2014). Evaluating single-case research data for systematic review: A commentary for the special issue. *Journal of School Psychology*, *52*, 129–133 (this issue).
Martinez, C. K., & Betz, A. M. (2013). Response interruption and redirection: Current research trends and clinical application. *Journal of Applied Behavior Analysis*, *46*, 549–554.
Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351.
McMillan, T. M. (2013). Outcome of rehabilitation for neurobehavioural disorders. *NeuroRehabilitation*, *32*, 791–801.
Moeyaert, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*, 83–103 (this issue).
Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of All Non-Overlapping Data (PAND): An alternative to PND. *Journal of Special Education*, *40*, 194–204.
Parker, R. I., & Vannest, K. J. (2012). Bottom-up analyses for single-case research designs. *Journal of Behavioral Education*, *21*, 254–265. http://dx.doi.org/10.1007/s10864-012-9153-1.
Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*, 303–322. http://dx.doi.org/10.1177/0145445511399147.
Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284–299.
Payne, S. W., & Dozier, C. L. (2013). Positive reinforcement as treatment for problem behavior maintained by negative reinforcement. *Journal of Applied Behavior Analysis*, *46*, 699–703.
Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*, 109–116.
R Development Core Team (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing (ISBN 3-900051-07-0, URL http://www.R-project.org/)
Rindskopf, D. M. (2014). Nonlinear bayesian analysis for single case designs. *Journal of School Psychology*, *52*, 71–81 (this issue).
Rindskopf, D., Shadish, W. R., & Hedges, L. V. (2012). *A simple effect size estimator for single-case designs using WinBUGS.* Washington D.C.: Society for Research on Educational Effectiveness.
Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York, NY: Henry Holt.
Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *Remedial and Special Education*, *8*, 43–48.
Scattone, D. (2008). Enhancing the conversation skills of a boy with Asperger's disorder through social stories and video monitoring. *Journal of Autism and Developmental Disorders*, *38*, 395–400.
Schertz, H. H., Reichow, B., Tan, P., Vaiouli, P., & Yildirim, E. (2012). Interventions for toddlers with autism spectrum disorders, an evaluation of research evidence. *Journal of Early Intervention*, *23*, 166–189.
Schreibman, L., Stahmer, A. C., Barlett, V. C., & Dufek, S. (2009). Brief report: Toward refinement of a predictive behavioral profile for treatment outcome in children with autism. *Research in Autism Spectrum Disorders*, *3*, 163–172 (SID = 9).
Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education*, *34*, 9–19. http://dx.doi.org/10.1177/0741932512440730.
Shadish, W. R. (1992). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 129–208). New York, NY: Sage.
Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K., Galindo, R., Nagler, E. D., et al. (2009). Using UnGraph® to extract data from image files: Verification of reliability and validity. *Behavior Research Methods*, *41*, 177–183.
Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*, 385–405. http://dx.doi.org/10.1037/a0032964.
Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2012). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*. http://dx.doi.org/10.3758/s13428-012-0282-1.
Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971–980. http://dx.doi.org/10.3758/s13428-011-0111-y (ERIC #ED530280).
Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*, 15–39 (this issue).
Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*, 41–70 (this issue).
Sherer, M. R., & Schreibman, L. (2005). Individual behavioral profiles and predictors of treatment effectiveness for children with autism. *Journal of Consulting and Clinical Psychology*, *73*, 525–538 (SID = 5).
Sidman, M. (1960). *Tactics of scientific research.* New York, NY: Basic Books.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis.* New York, NY: Oxford University Press.
Skinner, B. F. (1938). *The behavior of organisms.* New York, NY: Appleton-Century.
Skinner, B. F. (1972). A case history in scientific method. In B. F. Skinner (Eds.), *Cumulative record* (pp. 101–124). New York, NY: Appleton-Century-Crofts (Original work published 1956).
Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). *WinBUGS user manual.* version 2.0 (Retrieved from http://mathstat.helsinki.fi/openbugs/ManualsFrames.html)

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design Software Version 3.0.* (Available from www.wtgrantfoundation.org)

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2014). Analyzing longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods* (in press).

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, *52*, 105–122 (this issue).

Thorp, D. M., Stahmer, A. C., & Schreibman, L. (1995). The effects of sociodramatic play training on children with autism. *Journal of Autism and Developmental Disorders*, *25*, 265–282 (SID = 3).

van Oorsouw, W. M. W. J., Israel, M. L., von Heyn, R. E., & Duker, P. C. (2008). Side effects of contingent shock treatment. *Research in Developmental Disabilities*, *29*, 513–523. http://dx.doi.org/10.1016/j.ridd.2007.08.005.

Waldron, B., Casserly, L. M., & O'Sullivan, C. (2013). Cognitive behavioural therapy for depression and anxiety in adults with acquired brain injury. What works for whom? *Neuropsychological Rehabilitation*, *23*, 64–101.