Statistical Analyses of Single-Case Designs: The Shape of Things to Come

Author(s): William R. Shadish

Source: *Current Directions in Psychological Science*, April 2014, Vol. 23, No. 2 (April 2014), pp. 139-146

Published by: Sage Publications, Inc. on behalf of Association for Psychological Science

Stable URL: http://www.jstor.com/stable/44318737

Statistical Analyses of Single-Case Designs: The Shape of Things to Come

William R. Shadish
University of California, Merced

Abstract

Single-case-design researchers rarely used statistics in the past, but that is changing. In this article, I review the rapidly developing state of statistical analyses for single-case designs, including effect sizes, multilevel models, and Bayesian analyses. No analysis meets all the desiderata for an optimal single-case-design analysis, but this may be remedied in the near future. Single-case-design researchers will have incentives to use these analyses as they become more user-friendly and beneficial.

The single-case design (SCD) is one of the oldest and most respected design traditions for testing the effects of a treatment (Sidman, 1960; Skinner, 1938), especially in some areas of psychology and education (Shadish & Sullivan, 2011; Smith, 2011). A case is exposed to treatment, or not, sequentially in a time-series design, and a treatment effect is inferred if the outcome changes as predicted when treatment is present, but not otherwise. Figure 1 shows results from one example of an SCD (Lambert, Cartledge, Heward, & Lo, 2006). In this study, the treatment was aimed at reducing pupils' disruptive behaviors in two classrooms, and visual analysis of Figure 1 suggests that it worked (which SCD researchers call a *functional relationship*). Figure 1 shows a reversal (ABAB) design, but other variants of SCDs include the multiple-baseline design, the alternating-treatment design, the changing-criterion design, and designs that combine these four types (Shadish & Sullivan, 2011). In psychology and education, SCDs are used in such diverse fields as applied behavior analysis, autism, behavior modification and therapy, child and family therapy, developmental disorders, exceptional children, intellectual disabilities, language disorders, learning disabilities, neuropsychological rehabilitation, school psychology, special education, speech therapy, and sport and exercise psychology (Shadish & Sullivan, 2011). SCDs are also used in medicine, in which they are called *N-of-1 designs*, to study asthma, arthritis, fibromyalgia, medication choices, pain treatment, surgery, traumatic brain injury, and attention deficit hyperactivity disorder (Gabler, Duan, Vohra, & Kravitz, 2011).

Researchers, practitioners, and policymakers recognize the ability of SCDs to contribute to evidence-based practice reviews. The What Works Clearinghouse and the American Psychological Association's Division 16 Task Force on Evidence-Based Interventions in School Psychology have recommended acceptance of SCDs for evidence-based practice reviews (Kratochwill et al., 2010; Kratochwill & Stoiber, 2002). The U.S. Department of Education's National Center for Special Education Research allows SCDs to be used instead of randomized experiments for some efficacy studies. Oxford University's Centre for Evidence-Based Medicine, which works closely with the *British Journal of Medicine*, accepts N-of-1 trials as evidence about effects of medical interventions and ranks them at the same level as randomized experiments (Howick et al., 2011). Medical researchers are extending the Consolidated Standards of Reporting Trials (CONSORT) standards for reporting randomized experiments to reporting standards for N-of-1 trials (Shamseer et al., 2012).

However, an obstacle to full acceptance of SCDs among those outside the SCD community is the general failure to use statistics to analyze SCDs. Dating back to Skinner (1938), most SCD researchers have followed a visual-analysis tradition in which inferential statistics are

Corresponding Author:
William R. Shadish, Psychological Sciences, University of California,
5200 N. Lake Rd., Merced, CA 95343
E-mail: wshadish@ucmerced.edu

eschewed (Kratochwill, Levin, Horner, & Swoboda, in press)—despite the use of statistics having been proposed by members of the SCD community for almost as long. Some objections to using statistics are principled, especially those based on skepticism that statistics can capture the nuances an SCD researcher considers when judging whether a functional relationship exists (Parker & Vannest, 2012). More pragmatic reasons are habit, lack of advanced statistical curricula in SCD training programs, absence of consensus about which type of analysis is best, and a dearth of incentives to use statistics.

Yet this avoidance is changing because of (a) pressure from evidence-based-practice communities to speak the statistical language in common use; (b) expectations from granting agencies and journal editors for quantitative work in addition to visual analysis; (c) generational changes within the SCD communities; (d) greater collaboration between SCD researchers and statisticians that corrects misperceptions and builds common agreements about the role of statistics; (e) the development of a cadre of statisticians who are interested in the intellectual challenges of SCD analysis, who have ready access to virtually
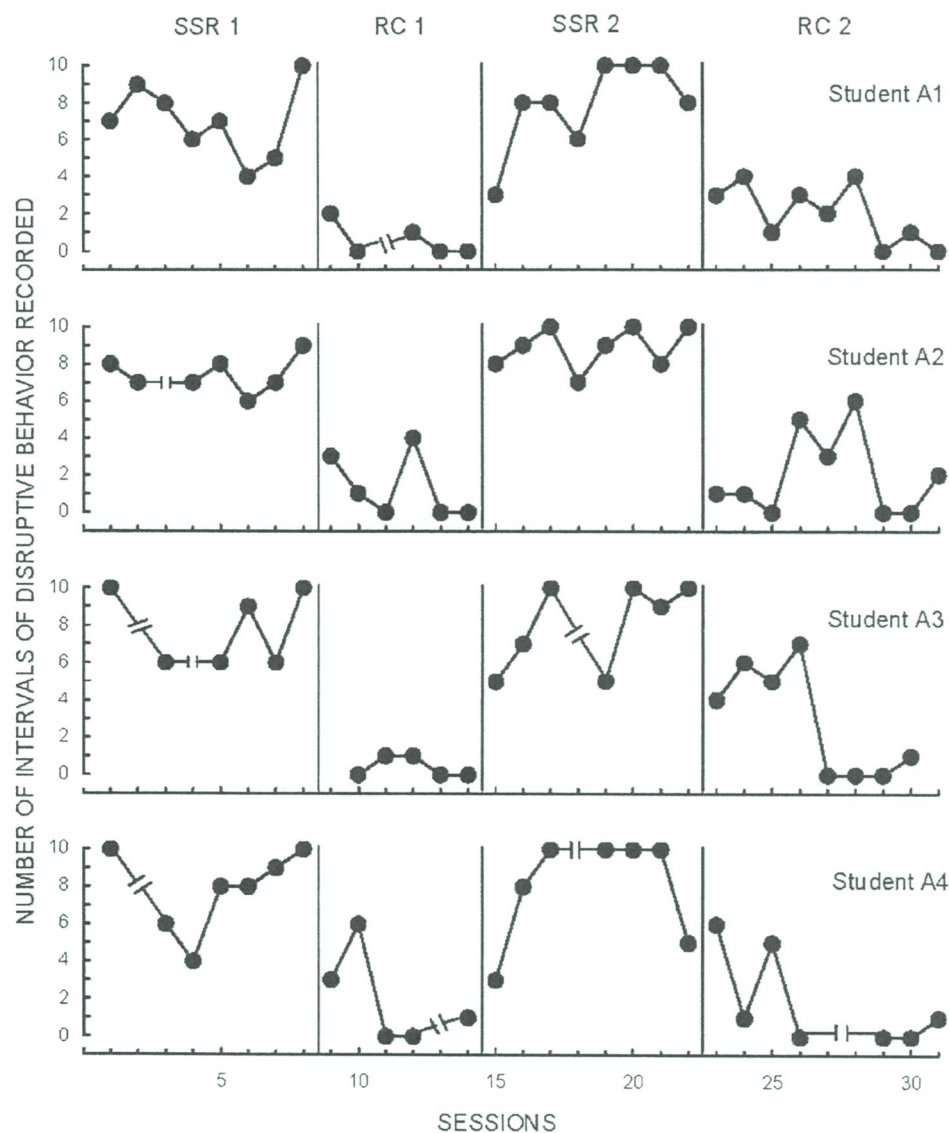

Fig. 1. Number of intervals of disruptive behavior recorded during single-student responding (SSR) and response-card (RC) conditions. Each row of graphs represents data for one student observed over at least 26 sessions. The SSR condition was a no-treatment condition, RC was the treatment condition, and both conditions were run twice to form an ABAB design. Successful treatment is indicated by a lower number of observations during RC than SSR. Reprinted from "Effects of Response Cards on Disruptive Behavior and Academic Responding During Math Lessons by Fourth-Grade Urban Students," by M. C. Lambert, G. Cartledge, W. L. Heward, and Y. Lo, 2006, *Journal of Positive Behavior Interventions, 8,* pp. 94–95. Copyright 2006 by Sage. Reprinted with permission.

Figure 1. (continued)

Fig. 1. (continued)
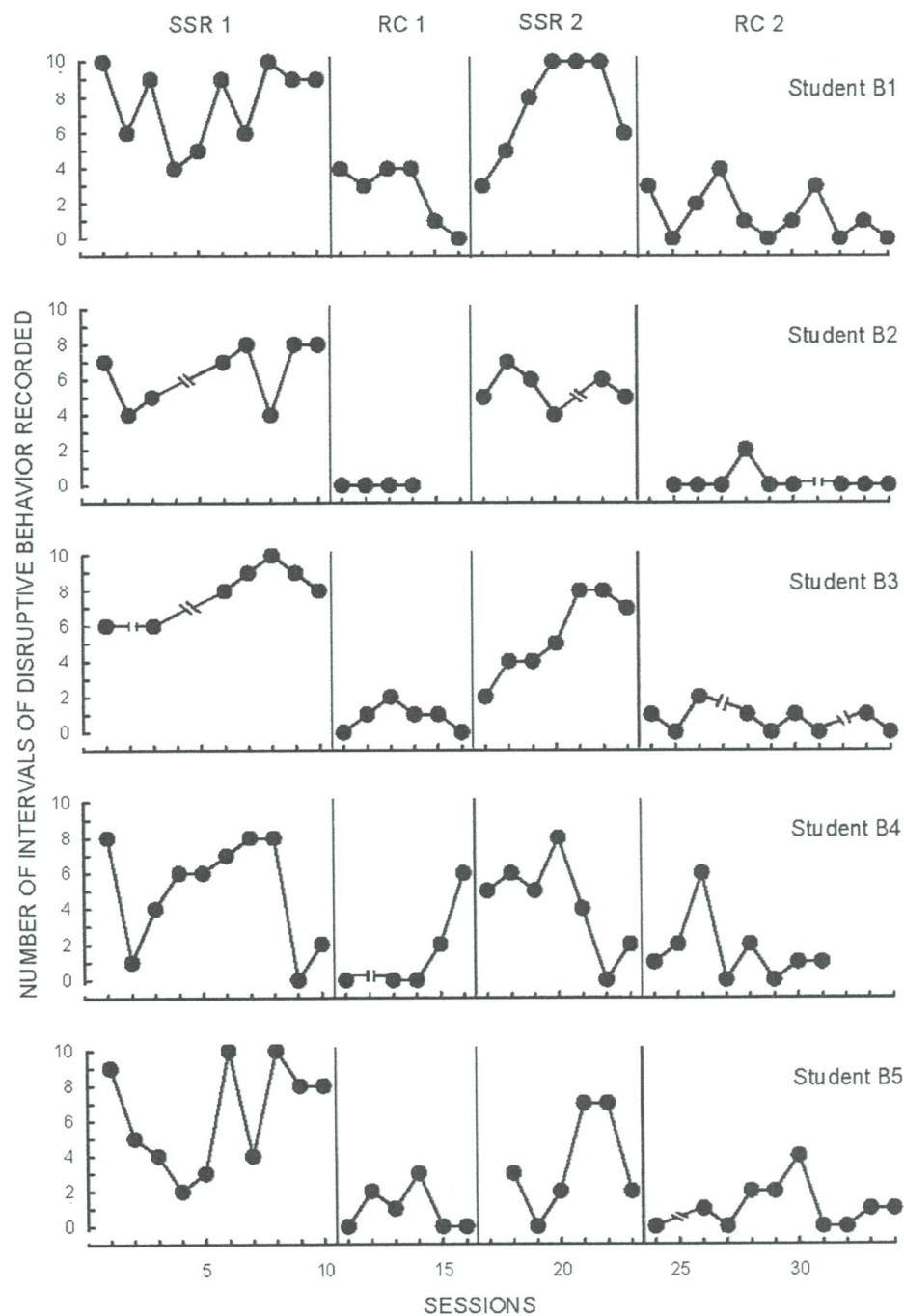
all published SCD data because they have the ability to digitize published graphs, and who are therefore already developing and conducting analyses and meta-analyses of SCD data (Shadish, in press-b); and (f)

increasing recommendations to use statistics by various entities providing standards for SCDs (Smith, 2011). Consequently, the past 10 years have seen great progress toward formally developed statistics capable of addressing

the distinctive needs of SCD research (Shadish, in press-a). In this article, I will review and evaluate the most recent and promising developments.

A Review of Current Developments

As much as past SCD statistics need critical review, I cannot provide that in this article. Instead, my focus is on current advances and the reasons for them.

Standardized effect sizes

Formally developed standardized effect sizes are of current interest because they tend to be the coin of the realm in evidence-based practice reviews, especially when meta-analysis is used to decide what works, or when results from SCDs are compared with results from between-groups studies, such as randomized experiments. The phrase "formally developed" means the effect sizes are derived analytically using statistical theory. A standardized effect size remains on the same measurement scale no matter what outcome measure is used. Examples are d statistics, odds ratios, and correlation coefficients. Because evidence-based practice reviews nearly always combine results over very different outcome measures, standardized effect sizes are needed.

Past proposals for a standardized effect size in SCD research are intriguing and intuitive. Overlap statistics are a good example (Parker, Vannest, & Davis, 2011). To judge whether a treatment worked, SCD researchers visually examine the overlap of treatment and baseline observations for each case (Kratochwill et al., in press). As shown in Figure 1, for example, the treatment used by Lambert et al. (2006) reduced students' disruptive behavior relative to baseline levels, resulting in little overlap of treatment and baseline observations. However, past SCD effect sizes lacked formal development from clear assumptions in statistical theory. As a result, their confidence intervals and significance tests are of unclear validity or are nonexistent, their power to detect effects is unknown, and we know little about how they perform in the face of variation in the number of observations per phase (*phase length*), in how observations change systematically over time even in the absence of treatment (*trend*), in how correlated observations within each case might be (*autocorrelation*), and in the outcome-measurement metrics (e.g., count, percentage, normally distributed data).

In comparison, consider the recent d statistic that Hedges, Pustejovsky, and Shadish (2012, 2013) developed. Its standard error is formally derived from clear distributional assumptions, it takes the autocorrelation into account, it has power analyses, and simulations have shown how it performs over a variety of circumstances common in SCD research. SPSS macros with graphical user interfaces (GUIs) are available to estimate this d and its power, and demonstrations show how to use the d meta-analytic statistics that cannot be applied with past SCD effect sizes (Shadish, Hedges, et al., 2013; Shadish, Hedges, & Pustejovsky, in press). It is explicitly derived from normal-distribution theory and explicitly assumes no trend, which are indeed limitations—but at least we know the kind of data to which it does apply! This work can be (and is being) extended to non-normally distributed data with trend. In contrast, past effect sizes are applied to all kinds of outcome distributions, to data with or without trend and with or without autocorrelation, with little thought to whether such applications are valid. The latter is, of course, remediable. Given how intuitive some past effect sizes are for SCD research, we hope that more formal statistical development of them occurs.

Analyses based on regression models

A second approach to the analysis of SCDs is regression or its extensions, such as multilevel models. Regression is formally developed statistically for many purposes relevant to SCD research, it can account for trend and autocorrelation, and it can assume non-normal outcome distributions that are appropriate for the count and rate measures that predominate in SCD research—although SCD researchers have rarely taken advantage of all these possibilities. Problems with past regression analyses are that they have been applied only within each case even though they could analyze all cases simultaneously, they do not easily provide statistical tests for whether cases differ significantly from each other, they usually assume trend is linear, and they do not typically produce a standardized effect-size estimate.

Multilevel models can easily address the first two of these problems (Moeyaert, Ferron, Beretvas, & Van den Noortgate, in press; Shadish, Kyse, & Rindskopf, 2013). For the data shown in Figure 1, past approaches would produce nine separate regression equations, one for each case, and stop there. Multilevel models would produce the same nine regression equations, but then extract and analyze the regression coefficients from each of these nine models (e.g., the nine regression coefficients for the treatment effect). So, they can test both the significance of the average of these coefficients (e.g., the average treatment effect over nine cases) and whether the nine regression coefficients differ significantly from each other (e.g., whether the treatment effect varies significantly over cases, although the power of that test can be low given the small number of cases often used in SCDs; see Shadish, Kyse, & Rindskopf, 2013).

Unfortunately, SCD researchers who use regression and multilevel models (both are parametric models) nearly

always assume that trend does not exist or is linear (always increasing or decreasing in a straight line) as opposed to nonlinear (as seen when, e.g., a delayed effect occurs but then affects the outcome rapidly and, finally, reaches a ceiling beyond which no further improvement occurs). If trend is not linear, regression coefficients and standard errors will be wrong if linearity is assumed. This is a major problem, because parametric methods require the researcher to correctly identify the form of the trend, but the researcher rarely knows that form. Recent semiparametric regression models (e.g., generalized additive models) help solve this problem. Semiparametric regression combines the usual parametric predictors (e.g., treatment) with nonparametric smoothing techniques (e.g., for trend) that allow the data to suggest the form of the trend. Examples suggest that trend in SCDS is often nonlinear, and that these models can be practical and useful for the analysis of SCDs (Shadish, Zuur, & Sullivan, 2014). Figure 2

is a graph taken from Shadish et al. (2014) illustrating results from an application of generalized additive models to the Lambert et al. (2006) data shown in Figure 1. The analysis allowed trend for each case to have a different linear or nonlinear functional form. For example, Case 6 has no significant trend, Case 2 has significant linear trend, and other cases have significant nonlinear trends that are very wiggly (e.g., Case 9).

Bayesian statistics

The most recent development is the application of Bayesian statistics to SCD data (Rindskopf, in press; Swaminathan, Rogers, & Horner, in press). Bayesian methods can produce more valid results in the presence of small sample sizes that are characteristic of SCDs (Shadish & Sullivan, 2011). Bayesian methods have other advantages when compared with the usual (frequentist)
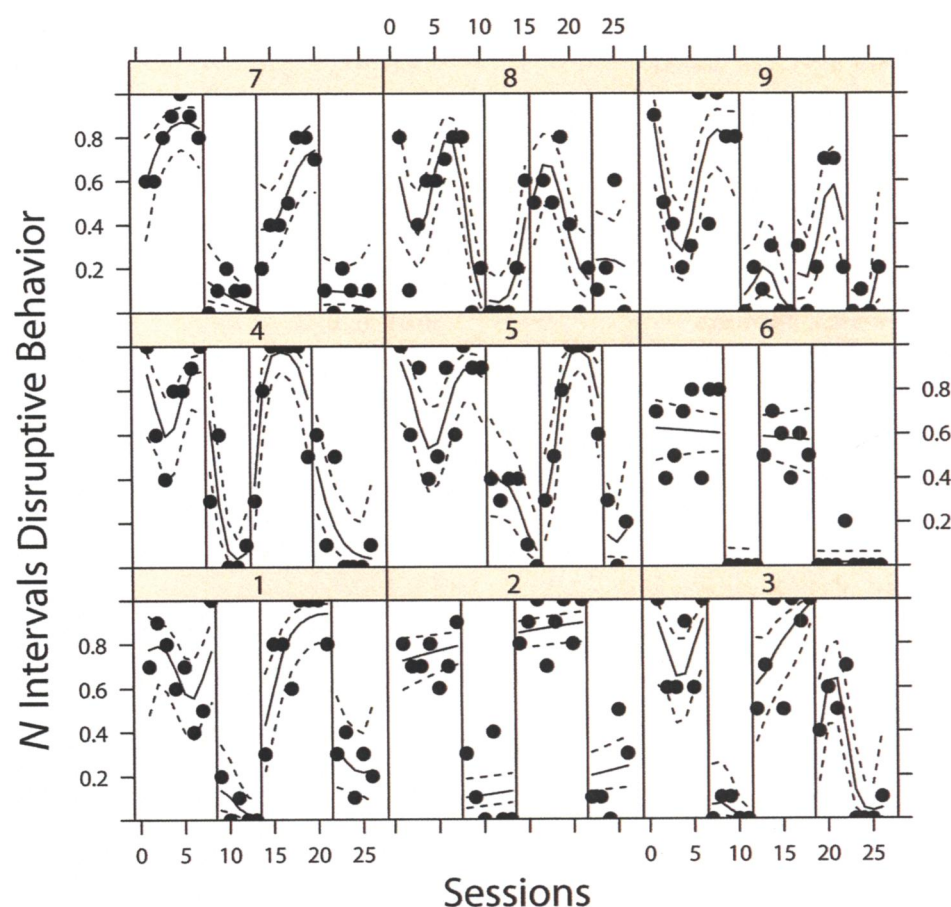

Fig. 2. A multipanel scatter plot of a generalized additive model semiparametric regression analysis in which trend varies as either linear or nonlinear for each case. Black dots are original raw data, solid lines are smoothed predicted values, and dashed lines are 95% confidence intervals around those predicted values. Data were drawn from Lambert, Cartledge, Heward, and Lo (2006); each of the nine plots represents a case shown in Figure 1 (1 = A1, 2 = A2, 3 = A3, 4 = A4, 5 = B1, 6 = B2, 7 = B3, 8 = B4, 9 = B5). Reprinted from "Using Generalized Additive (Mixed) Models to Analyze Single Case Designs," by W. R. Shadish, A. F. Zuur, and K. J. Sullivan, 2014, *Journal of School Psychology*, *52*, p. 160. Copyright 2014 by the American Psychological Association. Reprinted with permission.

analyses. For example, they take into account prior evidence and beliefs about parameters and can adjust those priors given data from a new SCD study. They provide more natural interpretations of results, although this is not obvious without careful comparison of Bayesian and frequentist interpretations. Rindskopf (in press) provides a readable introduction to Bayesian statistics as they relate to SCD data. Programs for Bayesian statistics have a steep learning curve, but the effort is worthwhile, and more user-friendly interfaces will appear (e.g., Woodward, 2011).

Swaminathan et al. (in press) have also proposed a *d* statistic that seems to produce estimates consistent with the more formally developed Hedges et al. (2012, 2013) *d* statistic. The same methods could be used to obtain a *d* statistic from multilevel and generalized additive models.

Criteria for the Ideal SCD Data Analysis

In prior work (Shadish, in press-b), I suggested that the ideal statistical analysis for SCD data would:

1. Accurately model trend.
2. Accurately model error structures with either auto-correlation or random effects.
3. Produce a standardized effect size suitable for meta-analysis.
4. Correctly model outcome variable distributions (e.g., normal, binomial, Poisson).
5. Be accompanied by appropriate power analyses.
6. Be accessible to SCD researchers through macros, GUIs, and examples with syntax.

The importance of the last criterion cannot be overstated. Many clinical scientists will understandably use simple statistical programs even if they are not state of the art (e.g., Bloom, Fischer, & Orme, 2009; Borckardt et al., 2007).

Table 1 shows how well the main analyses discussed above deal with each of the six criteria. No method meets all of the criteria. The cells in the table necessarily oversimplify, of course. For instance, some multilevel modeling programs have a point-and-click GUI, but most do not; power analyses exist for both regression and multilevel modeling, but they have not been specifically implemented for SCDs; and proposals for standardized effect sizes exist for some regression analyses, but not all.

Conclusions and Future Directions

A serious conceptual concern for both quantitative (Molenaar, 2004) and visual analysts (Perone, 1999; Sidman, 1960) is a tension between *nomothetic* and *idiographic* assumptions in statistics. Nomothetic analyses focus on generalizations over individuals; idiographic approaches look at the uniqueness of individuals (Cone, 1986). Some analyses discussed in this article are more nomothetic, assuming a common process over cases (e.g., Hedges et al.'s *d*; much multilevel modeling), but for some idiographic researchers, this assumption is at best an empirical question and at worst never useful. Yet some analyses (e.g., Hoeppner, Goodwin, Velicer, Mooney, & Hatsukami, 2008; Shadish et al., 2014) may meet both idiographic and nomothetic needs, and the trade-offs between those needs require more attention (Shadish, in press-b).

As this concern illustrates, in a field where many flowers currently bloom, who can predict the one on which we will finally settle our gaze? What I want, I think, is a user-friendly program dedicated to SCD research that has multilevel, semiparametric, and Bayesian capabilities; that produces a standardized effect size appropriate to

Table 1. A Sample of Current Approaches to the Analysis of Single-Case Designs and Their Capabilities

Approach	Model trend	Model error	Standardized effect size	Outcome distributions	Power analysis	Accessible user interface
Most overlap statistics	—	—	+	—	—	+
Tau-U overlap statistic	+	—	+	—	—	+
Most *d* statistics	?	?	+	—	—	—
Hedges et al. *d* statistic	—	+	+	—	+	+
Regression	+	+	?	+	—	?
Multilevel models	+	+	—	+	—	?
Generalized additive models	+	+	—	+	—	—

Note: A plus sign (+) means that the method clearly meets a given criterion. A question mark (?) means some implementations of the method may meet the criterion. The absence of either of these two, indicated by a dash, means the method generally does not meet the criterion.

the outcome distribution; and that produces publication-quality graphics (best of all, make it free!). That would meet or exceed all of my desiderata for an ideal analysis. Swaminathan et al. (in press) are working on a program with many such characteristics. Others will surely compete in that arena. More work needs to be completed before such a program can be realized fully, particularly the formal statistical development of effect sizes for non-normally distributed SCD data. Such work is in progress.

Others may want something else, and a special tension may exist between the needs of SCD clinicians and practitioners and those of SCD researchers. Furthermore, over the next few years, new methods may appear, or older methods may be resurrected and improved. For example, overlap statistics have great intuitive appeal, and researchers are working to overcome their limitations (Parker et al., 2011). Collaboration with statisticians might allow for their more formal statistical development. Similarly, traditional time-series analyses could be used to analyze SCDs (Gorsuch, 1983), but seemingly promising starts (e.g., Velicer, 1994) have languished for lack of further attention. Yet even with such uncertainties, the next 5 years will likely see an emerging consensus on an ideal analytic method for SCDs. We are already close to that goal.

Recommended Reading

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Methods for the design and analysis of single-case research.* Hillsdale, NJ: Erlbaum. A book with a good introduction to the design of single-case research and to some analyses extant at the time of publication.

Parker, R. I., & Vannest, K. J. (2012). (See References). An argument from a single-case-design researcher about why many statistics should not be applied to single-case-design data.

Parker, R. I, Vannest, K. J., & Davis, J. L. (2011). (See References). A review of overlap statistics.

Shadish, W.R. (Ed.). (in press-a). (See References). A special issue of a journal demonstrating a wide array of recently developed ways to analyze or meta-analyze single-case designs, with commentaries.

Declaration of Conflicting Interests

Funding

References

Bloom, B., Fischer, J., & Orme, J. (2009). *Evaluating practice: Guidelines for the accountable professional* (6th ed.). Boston, MA: Allyn & Bacon.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2007). Clinical practice as natural laboratory for psychotherapy research. *American Psychologist, 63*, 77–94.

Cone, J. D. (1986). Idiographic, nomothetic, and related perspectives in behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 111–128). New York, NY: Guilford.

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.

Gorsuch, R. (1983). Three methods for analyzing limited time series ($N = 1$) data. *Behavioral Assessment, 5*, 141–154.

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, 4*, 324–341. doi:10.1002/jrsm.1086

Hoeppner, B. B., Goodwin, M. S., Velicer, W. F., Mooney, M. E., & Hatsukami, D. K. (2008). Detecting longitudinal patterns of daily smoking following drastic cigarette reduction. *Addictive Behaviors, 33*, 623–639.

Howick, J., Chalmers, I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., . . . Thornton, H. (2011). *The 2011 Oxford CEBM Evidence Levels of Evidence* (Introductory Document). Oxford Centre for Evidence-Based Medicine. Retrieved from http://www.cebm.net/index.aspx?o=5653

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (in press). Visual analysis of single-case intervention research: Conceptual and methodological considerations. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances.* Washington, DC: American Psychological Association.

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the *Procedural and Coding Manual* of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17*, 341–389.

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, 8*, 88–99.

McLaughlin, J. E., & Ebanks, C. (2012, April). *Grant writing workshops for efficacy and replication projects and effectiveness projects* [PowerPoint presentation transcript]. Retrieved from http://ies.ed.gov/funding/webinars/pdf/April24efficacy_FY2013.pdf

Moeyaert, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (in press). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*, 201–218.

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analyses for single-case research designs. *Journal of Behavioral Education, 21*, 254–265. doi: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. doi:10.1177/0145445511399147

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*, 109–116.

Rindskopf, D. M. (in press). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*.

Shadish, W. R. (Ed.). (in press-a). Analysis and meta-analysis of single-case designs [Special Issue]. *Journal of School Psychology*.

Shadish, W. R. (Ed.). (in press-b). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (in press). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2013). A *d*-statistic for single-case designs that is equivalent to the usual between-groups *d*-statistic. *Neuropsychological Rehabilitation*. Advance online publication. doi:10.1080/09602011.2013.819021

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. doi:10.1037/a0032964

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. doi:10.3758/s13428-011-0111-y

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*, 149–178.

Shamseer, L., Sampson, M., Bukutu, C., Barrowman, N., Altman, D., Moher, D., & Vohra, S. (2012). P05.50. CONSORT extension for N-of-1 trials (CENT) guidelines. *BMC Complementary and Alternative Medicine, 12*(Suppl. 1), P410. Retrieved from http://www.biomedcentral.com/1472-6882/12/S1/P410

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms*. New York, NY: Appleton-Century.

Smith, J. D. (2011). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (in press). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*.

Velicer, W. F. (1994). Time series models of individual substance abusers. *NIDA Research Monographs, 142*, 264–301.

Woodward, P. (2011). *Bayesian analysis made simple: An Excel GUI for WinBUGS*. Boca Raton, FL: CRC Press.