Behav Res (2011) 43:971–980
DOI 10.3758/s13428-011-0111-y

Characteristics of single-case designs used to assess intervention effects in 2008

William R. Shadish · Kristynn J. Sullivan

Published online: 8 June 2011

Abstract This article reports the results of a study that located, digitized, and coded all 809 single-case designs appearing in 113 studies in the year 2008 in 21 journals in a variety of fields in psychology and education. Coded variables included the specific kind of design, number of cases per study, number of outcomes, data points and phases per case, and autocorrelations for each case. Although studies of the effects of interventions are a minority in these journals, within that category, single-case designs are used more frequently than randomized or nonrandomized experiments. The modal study uses a multiple-baseline design with 20 data points for each of three or four cases, where the aim of the intervention is to increase the frequency of a desired behavior; but these characteristics vary widely over studies. The average autocorrelation is near to but significantly different from zero; but autocorrelations are significantly heterogeneous. The results have implications for the contributions of single-case designs to evidence-based practice and suggest a number of future research directions.

Keywords Single-case design · Characteristics · Outcome studies

Single-case designs (SCDs) not only go by many names, but also come in many variants. For example, they have been called the intrasubject replication design (Gentile, Roden, & Klein, 1972), reversal design (Gentile et al., 1972), *n*-of-1 design (Center, Skiba, & Casey, 1986), intrasubject design (Center et al., 1986), intrasubject experimental design (White, Rusch, Kazdin, & Hartmann, 1989), individual organism research (Michael, 1974), $N = 1$ study (Strube, Gardner, & Hartmann, 1985), *N* of 1 data (Gorsuch, 1983), and one-subject experiment (Edgington, 1980). They include variations such as the ABAB design, alternating treatments design, multiple-baseline design, and changing criterion design (Hayes, 1981). In all variants, however, a case (often a person, but sometimes an aggregate unit such as a classroom) is measured on an outcome variable repeatedly over time. Observations occur during treatment phases, when an intervention is present, and during baseline and/or maintenance phases, when the intervention is not present. Baseline and treatment measurements are then compared to assess whether a functional relationship exists between the intervention and the outcome variable—whether the outcome changes either in level or in slope when the treatment is introduced but does not when it is absent.

Although SCDs are used in clinical and applied areas of psychology and education, little recent data exist to characterize all the variants that go under the rubric of SCDs (e.g., Kratochwill & Brady, 1978). Such data might include the kinds of SCDs used, the number of phases and data points in them, their frequency across a host of topic areas, and the serial dependency of errors (autocorrelation). These data could be useful for a variety of reasons. For example, they could be used to clarify the extent to which current SCDs meet extant standards for good SCDs or for evidence-based practice, such as the What Works Clearinghouse (WWC) Standards for the use of SCDs in evidence-based practice reviews (Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010). They could

W. R. Shadish · K. J. Sullivan
University of California,
Merced, USA

W. R. Shadish (✉)
School of Social Sciences, Humanities and Arts,
University of California,
Merced, 5200 North Lake Rd,
Merced, CA 95343, USA
e-mail: wshadish@ucmerced.edu

also inform statistical methods for analyzing single-case designs that make assumptions about any of these characteristics, such as the autocorrelation. Finally, they could be used to help in the design of simulation studies—the original purpose of the present study. The present study reports results from a survey of SCDs that provide such data.

Method

Procedure

Location of journals and studies Shadish and Rindskopf (2007) compiled a list of meta-analyses of SCDs. We included in our sample every journal that had published one of those meta-analyses. In addition, we scanned the reference lists of those meta-analyses to locate other journals that had published SCDs. The resulting list consisted of 22 journals. One of these journals (*Remedial Education*) did not exist in 2008, and so we included 21 journals in this review. We reviewed all issues of these 21 journals from the year 2008 for studies that contained an SCD. Table 1 lists the journals and some basic descriptive statistics about them.We limited this to the most recent year before the start of this project in order to reflect most recent practice and to keep the size of this very large project tractable. This is not a complete list of all journals that publish SCDs. In retrospect, others have called to our attention a number of other journals that publish SCDs, which we list here for the record (*American Journal of Speech, Language, Hearing; Behavior Therapy*; *Education and Treatment of Children*; *Child and Family Behavior Therapy*; *Journal of Behavioral Education*; *Journal of Early Intervention*; *Journal of Hearing, Speech, Language Research*; *Journal of Positive Behavior Interventions*; *Speech, Language, Hearing in the Schools*; *Topics in Early Childhood Special Education*; *Topics in Language Disorders*). Since some of these journals, such as the ones on language disorders, reflect entire topics that might not be represented in our sample, generalization from our results should be made with care.

Figure 1 shows a flow diagram of the study selection process. The 21 journals published in 2008 yielded 1,098 empirical articles—that is, studies that gathered data using any empirical methodology (Table 2). We classified each methodology as an assessment study (61%), an intervention study (26%), or a review (13%). Assessment studies could be surveys, comparisons of (nonmanipulable) groups, psychometric studies such as scale development, diagnostic studies, examinations of other characteristics of a sample, qualitative or case study reports, or description of program implementation. Intervention studies could be randomized experiments, nonrandomized experiments, or single-case designs. Reviews could be either narrative or quantitative (meta-analytic). Since some of these categories can overlap—for example, a case study concerning diagnosis—some of these distinctions are not completely orthogonal.

We read all 1,098 empirical articlesto locate SCDs. Studies used many different terms to identify SCDs: multiple-baseline designs, ABAB designs and their variants (in our cases, e.g., ABCD, AB, or ABABAB designs), alternating treatment designs (sometimes with a baseline), multielement designs, multiple-probe designs, concurrent-chain designs, reversal designs, within-person designs, changing criterion designs, and parallel treatment designs, as well as single-case designs (not otherwise specified). Regardless of what the authors called the SCDs, they were included in the study pool if (1) they measured an outcome more than once before the implementation of treatment and continued measuring the same outcome more than once after a treatment had been implemented across multiple data points on a single case; (2) they reported results separately for each case; and (3) the aim of the study was to assess the effects of an intervention. However, we did include cases with only one baseline or only one treatment assessment if they were used in the context of a set of cases where the first criterion was otherwise met, as in a multiple-baseline design. We excluded articles that did not use SCD methodologies. Finally, we excluded articles that gathered data on a single participant reported in an SCD format but that did not implement a treatment; for example, a few studies were concerned solely with methods for studying functional assessment and did not implement a treatment (e.g., Roscoe, Carreau, MacDonald, & Pence, 2008). This yielded 118 articles. A second graduate student then reviewed all of the journals and located eight SCDs that were overlooked in the first search, yielding a new total of 126 articles. In 4 articles, the title made it clear that they used SCDs; in 67 abstracts, using an SCD methodology was mentioned; and in 55 cases, the article text said that the study used an SCD methodology. In 13 of these 126 articles, data could not be digitized because the graphs were not sufficiently legible. They were excluded after requests to the authors for original data were not fruitful.[1] The final data set consisted of 113 studies.

[1] For other reasons, we have requested raw data from an additional dozen authors of SCDs. Over all requests we have ever made to such authors, the vast majority have never acknowledged receiving the multiple email requests, and none has ever sent the data. Only one ever provided a reason, that the data were owned by the school system where the research was done and so the author was not at liberty to release the data. Thus, the availability of the original raw SCD data for secondary analysis may itself be an interesting topic for future research.

Table 1 Descriptive statistics by journal

Journal Name	Studies with SCDs	Cases per Study	Dependent Variables per Case	Data Points per Case	Phases per Case
J. of Applied Behavior Analysis	37	3.46	2.08	28.98	10.60
Behavior Modification	16	3.94	1.69	38.47	4.18
Research in ASDs	12	3.50	3.42	35.74	9.03
Focus on Autism and other DDs	11	3.09	3.09	19.33	5.21
Research in DDs	9	3.56	2.44	29.39	12.98
J. of Autism and DDs	8	2.75	2.75	24.03	4.23
J. of Special Education	4	5.00	2.00	9.42	4.79
J. of School Psychology	3	3.33	1.67	15.39	2.50
Autism	2	6.50	2.00	16.81	4.95
J. of Emotional and BehaviorDisorders	2	1.50	4.00	24.20	2.60
Psychology in the Schools	2	7.00	2.00	19.50	2.20
Remedial and Special Education	2	4.50	2.00	17.11	3.67
School Psychology Review	2	2.50	2.00	38.90	6.70
Behavior Research and Therapy	1	9.00	1.00	8.67	7.22
Exceptional Children	1	4.00	2.00	12.50	4.00
School Psychology Quarterly	1	3.00	3.00	23.33	3.00
Correlations Among Characteristics					
Cases	−.226				
Dependent variables	.035	−.674**			
Data points	.444	−.534*	.321		
Phases	.550*	−.002	−.032	.374	

Note. J. = journal, DDs = developmental disabilities, ASDs = autism spectrum disorders

Coding

For nearly all studies, we digitized data points from electronic copies of graphs in the publications, using the computer program Ungraph®, which procedure yields data of extremely high reliability and validity (Shadish, Brasil, Illingworth, White, Galindo, Nagler, & Rindskopf, 2009). Once, we used original raw data the authors had included in a publication. Ten trained undergraduate students performed the initial data extraction. Each extraction consisted of two coordinates for each point in the graph. One was the session number, typically as reported on the horizontal axis of the graph; the other was the value of the dependent variable, typically as reported on the *y*-axis. After the initial digitization, the second author checked each extraction and fixed mistakes, appealing to the first author in difficult cases until 100% agreement was reached. Consistent with the high reliability reported in Shadish et al. (2009), nearly all mistakes were trivial rounding errors, either failing to round a discrete variable, such as a coordinate that should be an integer (e.g., a session number), or rounding continuous variables to integers when they should have included decimal points (e.g., a percent). Errors were more frequent in complicated graphs.

We assigned ID numbers to each journal, study, dependent variable, and case. We coded the type of design by adapting a typology presented in the WWC Standards for SCDs (Kratochwill et al., 2010): (1) phase change with reversal, which are designs in which a later phase reverts to a condition given in a previous phase, such as ABA or BCBC; (2) phase change without reversal, which are designs in which conditions change over phases but never revert to a condition from a previous phase, such as AB or ABCD; (3) multiple-baseline designs; (4) alternating treatment designs; (5) changing criterion designs; or (6) some combination of the previous five basic design types. These categories are not completely orthogonal. For instance, multiple-baseline designs typically include a phase change but have characteristics in addition to the two phase change categories and, so, can be coded separately.

We coded the direction of the dependent variable (i.e., if the treatment works, will the dependent variable increase or decrease) and the dependent variable metric (were the data reported as a count total, an average count, a percent based on a count, or a continuous or quasi-continuous variable). For the count data, we recorded whether the number of response opportunities

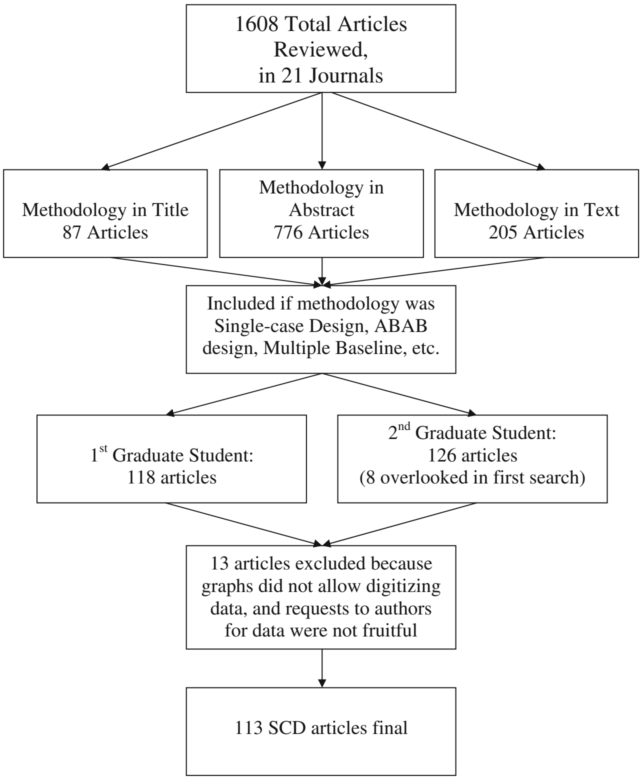
1608 Total Articles Reviewed, in 21 Journals
Methodology in Title 87 Articles
Methodology in Abstract 776 Articles
Methodology in Text 205 Articles
Included if methodology was Single-case Design, ABAB design, Multiple Baseline, etc.
1st Graduate Student: 118 articles
2nd Graduate Student: 126 articles (8 overlooked in first search)
13 articles excluded because graphs did not allow digitizing data, and requests to authors for data were not fruitful
113 SCD articles final

Fig. 1 Flow chart of study selection

was the same or different within case for every time point. We coded each digitized data point for the following variables: (1) type of phase (baseline, treatment, or maintenance/generalization); (2) phase number in order(e.g., first phase, second phase, third phase, and so forth), regardless of phase type; and (3) phase number in order separately for each type of phase (e.g., first baseline, second baseline, and so forth).

Results

Number of SCDs

Of the 1,098 empirical articles in these 21 journals, 113 (10.3%) reported results from 411 cases in which SCDs were used to assess the effects of an intervention. SCDs were used more often than either randomized or nonrandomized experiments, making them the most common method for assessing the effects of an intervention in these journals. Moreover, cases were often assessed on more than one outcome. So counting each unique combination of case and dependent variable, the total was 809 separate SCDs across the 113 studies. Thus, this data set may be one of the largest of its kind ever assembled.

Designs

More than half (54.3%) of the 809 SCDs used a multiple-baseline design (Table 3). Designs using a phase change with a reversal (8.2%) and alternating treatments designs (8.0%) were not uncommon. A substantial minority (26.1%) used designs best described as some combination of two or three of the five basic design types; for instance, 9.9% of SCDs were some combination of a multiple-baseline and an alternating treatments design. In general, designs were the same within a study over cases; for example, if a study used a phase change design with a reversal for one case, it nearly always used that same design for all cases in the study.

Cases

The name *single-case design* has often been read by those less familiar with this literature as meaning that each study reports results for only a single case. That perspective is mostly not an accurate characterization. The number of cases per study ranged from 1 to 13 and averaged 3.64 per study. The median and modal study reported results from 3 cases, and most studies (73.5%) had at least that many cases.

Outcome variables

Although the modal study measured results on one (N = 45; 39.8%) dependent variable per case, the majority (60.2%) measured two (N = 29; 25.7%), three (N = 22; 19.5%), or more (four to nine) dependent variables for each case. In the clear majority of the 266 instances of outcome measures (N = 192; 72.2%), researchers were trying to increase a desired behavior if the treatment was successful; in the rest (N = 74; 27.8%), researchers were trying to decrease an undesirable behavior. Finally, the metrics used in these measurements were nearly all (92.9%) some form of a count; the remainder were most often measures of time, such as time-on-task.

Number of data points

SCDs are, indeed, time series that are short, as compared with the usual 50–100 data points that the standard time series literature recommends are needed for statistical analysis—for example, to get adequate model identification in an ARIMA model (Box, Jenkins, & Reinsel, 1994; Velicer & Harrop, 1983). The median and modal number of data points in these 809 SCDs is 20, the range is 2–160, and 90.6% have 49 or fewer data points (Fig. 2). Some of the cases with few data points are somewhat misleading when taken individually. For instance, the case with 2 data points was part of a multiple-baseline design where the other cases

Table 2 Methods used in empirical articles by journal

	AJMR	AUT	BM	BRT	EC	FADD	JABA	JADD	JEBD	JSP	JIDR	JLD	JSP	JSE	LDRP	LDQ	PS	RSE	RASD	RDD	SPQ	SPR	Total	(%)
Assessment																								
Survey	11	8	3	44	5	5	2	42	1	22	41	14	22	2	3	5	22	11	10	10	15	21	319	(29%)
Group comparison	12	5	2	21	1	1	0	89	2	0	15	14	0	0	3	0	2	1	19	17	1	0	205	(19%)
Psychometrics	3	4	4	4	0	1	0	22	2	2	5	0	2	0	0	1	4	2	6	4	3	3	72	(7%)
Diagnostics	3	0	0	2	0	0	2	8	1	0	0	0	0	0	0	0	0	0	0	0	0	1	17	(2%)
Characteristics	3	4	1	1	4	0	0	8	6	0	5	2	0	0	1	0	0	1	0	0	0	1	37	(3%)
Qualitative/case study	0	3	0	0	2	2	0	5	0	0	2	0	0	2	0	0	3	2	0	0	0	0	21	(2%)
Program implementation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	(.1%)
Intervention																								
Randomized experiment	1	0	6	26	1	0	8	3	1	3	1	3	3	2	0	0	4	2	1	3	4	0	72	(7%)
Nonrandomized experiment	1	8	3	25	2	0	3	4	1	0	8	5	0	0	2	3	3	2	2	4	1	0	77	(7%)
Single-case design[1]	0	2	17	1	1	11	45	9	2	2	1	0	2	4	0	0	2	2	12	12	1	2	128	(12%)
Review																								
Narrative	0	0	11	4	6	1	4	10	1	2	16	0	2	8	8	3	29	5	10	1	13	3	137	(12%)
Quantitative	0	0	3	2	1	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	1	1	12	(1%)
Total	34	34	50	130	23	21	64	200	19	31	94	38	31	18	18	13	70	28	60	51	39	32	1098	(100%)

Note. AJMR = *American Journal of Mental Retardation*; AUT = *Autism*; BM = *Behavior Modification*; BRT = *Behaviour Research and Therapy*; EC = *Exceptional Child*; FADD = *Focus on Autism and Other Developmental Disabilities*; JABA = *Journal of Applied Behavior Analysis*; JADD = *Journal of Autism and Developmental Disabilities*; JEBD = *Journal of Emotional and Behavioral Disorders*; JIDR = *Journal of Intellectual Disability Research*; JLD = *Journal of Learning Disabilities*; JSP = *Journal of School Psychology*; JSE = *Journal of Special Education*; LDRP = *Learning Disabilities Research and Practice*; LDQ = *Learning Disability Quarterly*; PS = *Psychology in the Schools*; RSE = *Remedial and Special Education*; RASD = *Research in Autism Spectrum Disorder*; RDD = *Research in Developmental Disabilities*; SPQ = *School Psychology Quarterly*; SPR = *School Psychology Review.*

[1] Unlike in Table 1, in this table, single-case design includes both treatment studies and those without treatment but in which a single-case design format was used, such as a study only of functional analysis.

had more data points (Dancho, Thompson, & Rhoades, 2008). The study with 160 data points was also part of a multiple-baseline design in which victims of sexual assault were treated for PTSD and assessed for daily distress levels for nearly a year. Only 160 data points were extracted because the graph was quite compressed, so that making finer distinctions between days was not possible when the data were digitized.

The number of data points in the first baseline phase is of particular interest. Despite recommendations to delay the introduction of intervention until stable, flat baseline data have been obtained, there is often pressure to initiate

Table 3 Descriptive characteristics of SCDs

	Frequency	Percentage
Design Type		
Phase change with reversal (PCR)	66	8.2
Phase change without reversal (PCNoR)	7	0.9
Multiple baseline (MB)	439	54.3
Alternating treatments (AT)	65	8.0
Changing criterion (CC)	21	2.6
PCR + MB	97	12.0
PCR + AT	2	0.2
MB + AT	80	9.9
MB + CC	3	0.4
AT + CC	6	0.7
PCR + MB + AT	7	0.9
PCR + MB + CC	10	1.2
MB + AT + CC	6	0.7
Number of Cases Per Study		
1	19	16.8
2	11	9.7
3	37	32.7
4	19	16.8
5	8	7.1
6	10	8.8
8	3	2.7
9	4	3.5
10	1	0.9
13	1	0.9
Number of Dependent Variables Per Case		
1	45	39.8
2	29	25.7
3	22	19.5
4	5	4.4
5	4	3.5
6	4	3.5
8	3	2.7
9	1	0.9
Dependent Variable Metrics		
Count total, same[a]	6	2.2
Count percentage, same	76	28.5
Count mean, same	1	0.4
Count total, different	129	48.3
Count percentage, different	31	11.6
Count mean, different	5	1.9
Continuous	13	4.9
Quasi-continuous	6	2.2

[a] Same = number of trials at each time point is the same for every time point; Different = number of trials at each time pint is different for every time point

intervention early. For this reason, baselines are often shorter than would be desired. For example, the WWC Standards (Kratochwill et al., 2010) recommend a minimum of five data points in such a phase. This was not the case in many of the SCDs in our data set. Excluding alternating treatment designs where the design tends to preclude multiple points per phase, $N = 309$ (54.7%) of the SCDs had five or more points in the first baseline; $N = 24$ (4%) had one point, 37 (6.1%) had two points, 121 (20.1%) had three points, and 91 (15.1%) had four points in the first baseline phase. Some researchers would question whether short first baselines are sufficient to establish a stable baseline, and short baselines also reduce the amount of data available for reliable effect size calculations comparing baseline with treatment phases.

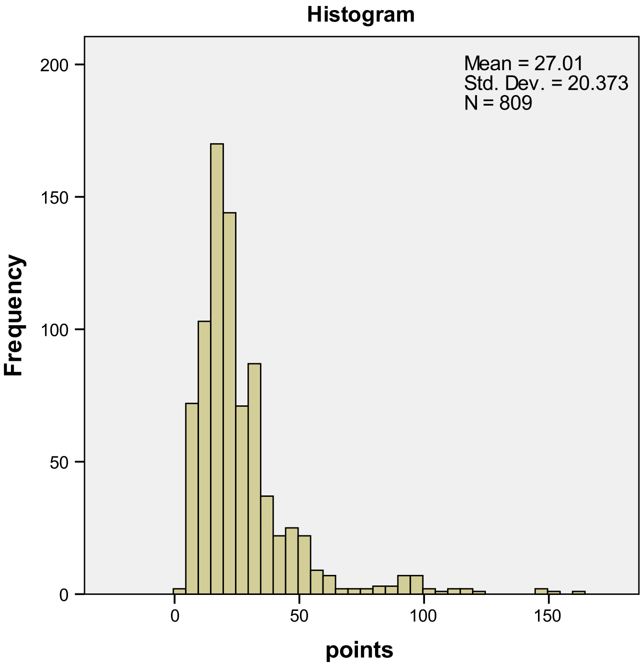

Fig. 2 Histogram of number of data points per case

Number of phases

We operationally defined a phase as an experimental condition and a change of phase as any contiguous change in experimental condition. The number of phases, so defined, in each SCD ranged from 1 to 98 (Fig. 3). This assessment includes probe, maintenance, or follow-up phases. The median number of phases was 4, and the mode was 2; and 78.6% of SCDs had 6 or fewer phases. Some unusual instances did occur. For example, eight SCDs had only 1 phase. One such study (Kurtz, Chin, Rush, & Dixon, 2008) examined two dependent variables in an AB design, measuring one of them during both phases but the other only during the treatment (B) phase. Consequently, while the overall design had 2 phases, the design for that particular dependent variable had only 1 phase. Conversely, large number of phases occurred in alternating treatment

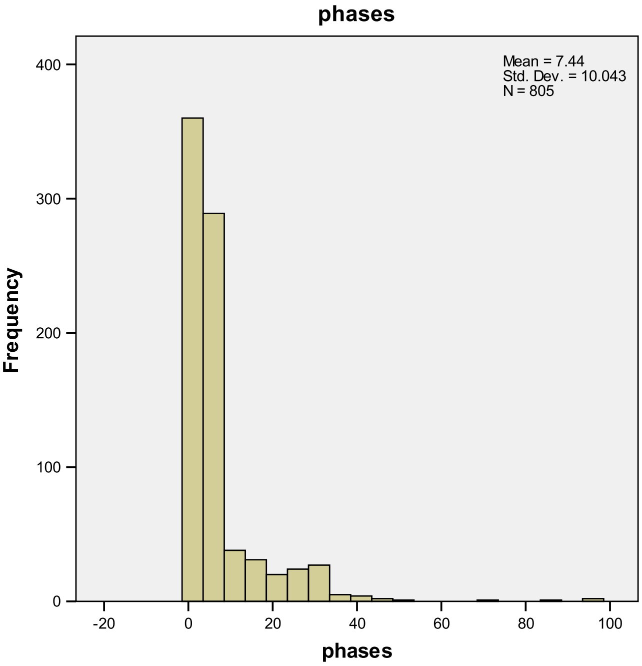

Fig. 3 Histogram of number of phases per case

designs. For instance, DeQuinzio, Townsend, and Poulson (2008) alternated training and probe trials within each of 49 sessions, resulting in the 98 phases described above as the high end of the range.

We also separately assessed the number of baseline, treatment, and maintenance/follow-up phases within each case. The distributions were positively skewed for all three, very similar to that in Fig. 3. The number of baseline phases ranged from 0 to 63, with a median and mode of 1. The explanations for few or many baseline phases are similar to those in the previous paragraph; for example, 63 baseline phases occurred in an alternating treatment design with a return to baseline after each treatment presentation. The number of treatment phases similarly ranged from 0 to 53, with a median of 2 and a mode of 1. Having no treatment for a particular case reminds us that all cases have to be considered in light of the overall design across all cases in a study. For example, Stephens (2008) used a multiple-probe design across three outcomes and four children, with an ABAB reversal for one child. The aim was to see whether training one child would result in imitation by the other children, so no training was given to the other children. The number of maintenance phases ranged from 0 to 10, with a median and mode of 0. The majority (68.0%) of SCDs had no maintenance or follow-up phases, and 98.4% had 2 or fewer such phases.

Differences among journals

Table 1 presents some of these basic characteristics separately for each journal. Journals differed significantly from each other for number of data points per case, $F(15, 793) = 10.073$, $p < .001$, and number of phases per case, $F(15, 793) = 6.645$, $p < .001$, but not number of cases, $F(15, 97) = 1.467$, $p = .133$, or number of dependent variables, $F(15,97) = 0.929$, $p = .535$. We do not report posthoc follow-up tests, because the small sample sizes in so many of the journals that published very few SCDs resulted in very low power. Table 1 also shows that journals that published more SCDs had designs with more phases and that studies with more cases had fewer dependent variables and data points. The latter result suggests a logistical trade-off that is not uncommon in the SCD literature. That is, you can have lots of cases with few data points or lots of data points on few cases, but it is rare to have both lots of cases and lots of data points within cases.

Autocorrelations

We calculated autocorrelations of the lag one residuals for each time series. First, we computed a linear regression of data points on time, treatment, and the interaction of the two. Treatment was a dummy code for treatment points versus baseline points (maintenance points were excluded from this and most subsequent analyses). A few SCDs ($N = 10$) were excluded from this analysis because they either had too few time points to fit residuals or had zero residual variance in the denominator of the autocorrelation. We calculated the lag one autocorrelation (r_j) using a standard estimator:

$$r_j = \frac{\sum_{t=1}^{n-1} y_t y_{t+1}}{\sum_{t=1}^{n} y_t{}^2} \qquad (1)$$

where y_t is the residual of the observation at time t_j and y_{t+1} is the residual at time $t_j + 1$ (Huitema & McKean, 1994). Figure 4 displays a histogram of the autocorrelations. They range from $r_j = -.931$ to .786.

While we could have taken the simple arithmetic average of these autocorrelations, doing so would not take sampling error into account. That is, autocorrelations from designs with many data points are more precise estimates of the population parameter than are those from short designs. Hence, we used a meta-analytic mean weighted by the inverse of the sampling variance of the autocorrelation:

$$v_j = \left(1 - \rho_j^2\right) / \left(t_j - 3\right) \qquad (2)$$

where ρ_j is the autocorrelation of the j^{th} case ($j = 1. . . k$), and t_j is the number of time points in the j^{th} case (Anderson, 1971). The random effects meta-analytic mean of these autocorrelations was $\bar{r}_j = -0.037$, very small but significantly different from zero given the large sample

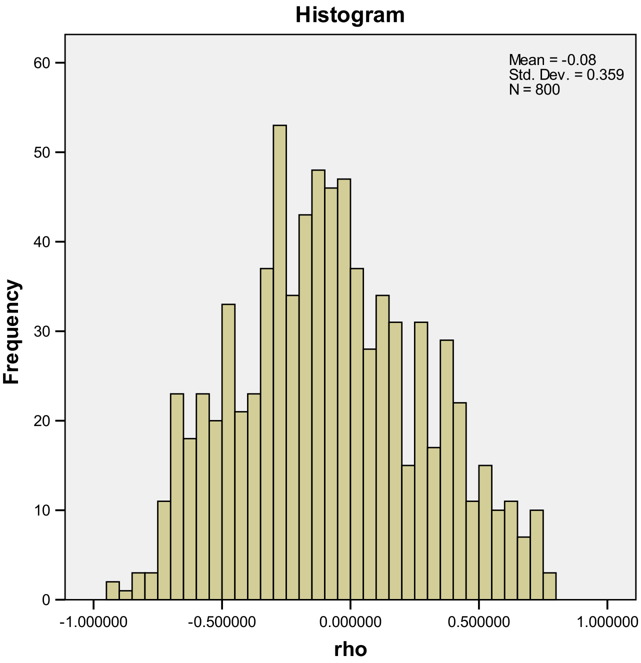

Fig. 4 Histogram of SCD autocorrelations

size (SE = 0.016, df = 798, p = .025). The variance component τ^2 = .155 was significant, Q = 4,306.18, df = 798, p < .001. The I^2 = 81% indicates that most of the variance in autocorrelations may be due to some systematic factors, rather than chance, such as the underlying stability of the dependent variable or the presence of floor or ceiling effects. For example, autocorrelations differ by the type of SCD used in our data set (Table 4; Q_b = 663.79, df = 4, p < .001). Alternating treatment and changing criterion designs had an average autocorrelation significantly less than zero, and the other designs had autocorrelations significantly higher than zero.

However, past authors have noted that the autocorrelation is negatively biased when the number of time points is small (Huitema & McKean, 1994; McKnight, McKean, & Huitema, 2000). The bias is approximately $-(1 + 3\rho_j)/t$ when the data are generated from a mean-only model and $-(P + 3\rho_j)/t$, where P is the number of parameters used to generate the data, when a more complex model is used, as is the case in the present example, where four parameters were used (Ferron, 2002). Figure 5 shows a scatterplot of the relationship between the autocorrelation (rho) and the number of time points (t) for each time series. A lowess smoother line is fit to the plot, and the line clearly shows a small drop in the size of the autocorrelations between 50 and 20 time points and a sharp drop below 20 time points. This would be consistent with the negative bias in small time series.

Given this bias, the observed autocorrelation (r) underestimates the population autocorrelation (ρ) by

$$r = \rho - \frac{P + 3\rho}{t}.$$

Since P = 4 in this study, a little algebra suggests a correction for both biases (Huitema & McKean, 2000):

$$\rho = -\frac{rt + 4}{t - 3}.$$

It is possible for this correction to result in $\rho > 1.00$. This happened in 18 cases that we set to $\rho = 1.00$. Using this correction, the random effects meta-analytic mean of these autocorrelations was $\bar{r}_j = 0.20$ (SE = .018, p < .001). The variance component τ^2 = .20 was also significant, Q = 9309.68, df = 780, p < .0001. If we repeat the test for whether this corrected autocorrelation varies by design type, the result is again significant (Q_b = 593.32, df = 4, p < .001; Table 4). This time, however, the autocorrelations for alternating treatment designs were not significantly different from zero, while all the other designs remained significantly different and in the same direction as before.

We caution readers that this correction must be viewed as approximate, not statistically exact. The autocorrelations exceeding unity suggest possible overcorrection, and the very large increase in the heterogeneity statistic is troubling. Past corrections for bias (e.g., Huitema & McKean, 1994) noted that success in reducing bias came at the cost of increasing the variance of the autocorrelations. Our results support that concern. Still, this analysis suggests that the meta-analytic average of the uncorrected autocorrelations may underestimate the true autocorrelation.

Table 4 Autocorrelations for different kinds of single-case designs

	Autocorrelation	Corrected Autocorrelation
Phase change with reversal (N = 64)	.088***	.191***
Phase change without reversal (N = 7)	.675***	.752***
Multiple baseline (N = 531)	.145***	.320***
Alternating treatments (N = 151)	−.179***	−.010
Changing criterion (N = 46)	−.085**	−.097**

** p < .01, ***p < .0001

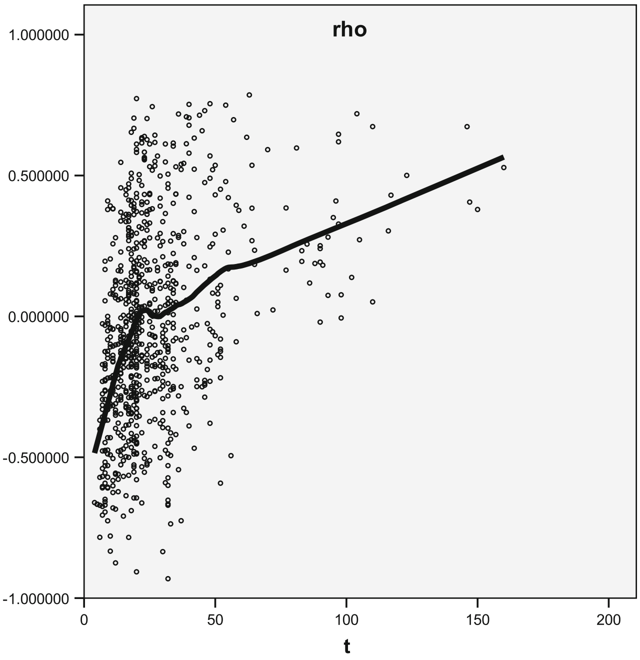

Fig. 5 Scatterplot of the autocorrelation (rho) against the number of time points (*t*)

Discussion

The original purpose of this study was to provide an empirical basis for setting the levels of the independent variables in a computer simulation for a new d-statistic proposed for use in SCDs (Shadish, Sullivan, Hedges, & Rindskopf, 2010). That statistic depends in part on all the facets we presented here, such as the number of cases, points, phases, and autocorrelations, so we needed to know the range and modal instances of these variables. Yet, after analyzing the data for that purpose, we realized that the data might be of wider interest. In part, this is simply due to its descriptive characterization of the literature, so that other SCD researchers can make more empirically supported statements when describing that literature. For instance, the modal study is a multiple-baseline design with 20 data points for each of three or four cases, where the aim of the intervention was to increase the frequency of a desired behavior.

Of particular interest is the fact that nearly all outcome variables were some form of a count. Most parametric statistical procedures assume that the outcome variable is normally distributed. Counts are unlikely to meet that assumption and, instead, may require other distributional assumptions. In some cases, for example, the outcome is a simple count of the number of behaviors emitted in a session of a fixed length, which has a Poisson distribution. In other cases, the result might be reported as a proportion: a count of successful trials divided by the total number of trials, where the number of trials is fixed and known over all sessions. Because this dependent variable is a proportion from a fixed number of binary (0, 1) observations, a binomial distribution may be appropriate. There are many such variations here, and this topic has received too little attention in the literature on statistical analysis of SCDs.

Yet the data are also interesting for the future research that they suggest.The data clearly make the case that SCDs are a commonly used methodology in areas such as behavior analysis, education, and developmental disabilities, more common than randomized or nonrandomized experiments. This finding is particularly significant given that SCDs have long been excluded from discussions of evidence-based practice, which traditionally have been based on randomized (and sometimes nonrandomized) controlled trials. Recent efforts to include SCDs in such reviews, like the WWC standards for the use of SCDs in evidence-based practice reviews (Kratochwill et al., 2010), can only be strengthened by empirical data supporting the contention that a large portion of the outcome literature will otherwise be missed.

These data also suggest the possibility that a majority of SCDs might fail to meet the evidence standards promulgated by, for example, WWC. For instance, WWC recommends five baseline datapoints before introducing an intervention, and it is apparent from the data that this will not be met for a substantial minority of cases. Similarly, WWC recommends three opportunities to demonstrate an effect, either by reversals or in a multiple-baseline design or similar context, and our impression of these data again suggests that that is not the case for a substantial minority of cases. Since the WWC criteria are cumulative—you must meet all the criteria, not just one—the chance of a majority of studies failing to meet them is increased. We do not yet have detailed data on this possibility but are conducting such a study now.

The autocorrelation data also raise questions. On the one hand, these data make clear that one cannot assume that the autocorrelation problem is negligible in the statistical analysis of SCDs (Huitema & McKean, 1994; McKnight et al., 2000). On the other hand, because the autocorrelations were statistically heterogeneous, we have much to learn about the conditions under which autocorrelations are strong or weak, positive or negative, or negligible. A first step toward that learning is to conduct a more detailed meta-analytic inquiry to identify predictors of heterogeneity such as the presence of floor and ceiling effects, the changeability of the dependent variable, and perhaps even some statistical aspects such as the method of computation of the autocorrelation. We are currently conducting such a meta-analysis now.

As evidenced by the willingness to include SCDs in evidence-based practice reviews by such organizations as the WWC, both researchers and policymakers are giving

SCDs increased attention. Standards are being developed, and new statistical analyses suggested (Shadish, Rindskopf, & Hedges, 2008). We hope studies like the present one will contribute to the accuracy of all that attention.

Author Note William R. Shadish and Kristynn J. Sullivan, School of Social Sciences, Humanities and Arts, University of California. 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.

References

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs: Prentice Hall.

Center, B. A., Skiba, R. J., & Casey, A. (1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387–400.

Dancho, K. A., Thompson, R. H., & Rhoades, M. M. (2008). Teaching preschool children to avoid poison hazards. *Journal of Applied Behavior Analysis, 41,* 267–271. doi:10.1901/jaba.2008.41-267

DeQuinzio, J. A., Townsend, D. B., & Poulson, C. L. (2008). The effects of forward chaining and contingent social interaction on the acquisition of complex sharing responses by children with autism. *Research in Autism Spectrum Disorders, 2,* 264–275. doi:10.1016/j.rasd.2007.06.006

Edgington, E. S. (1980). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2,* 19–28.

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers, 34,* 324–331.

Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5,* 193–198.

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

Hayes, S. C. (1981). Single-case experimental designs and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49,* 193–211.

Huitema, B. E., & McKean, J. W. (1994). Two biased-reduced autocorrelation estimators: r_{F1} and r_{F2}. *Perceptual and Motor Skills, 78,* 323–330.

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports, 87,* 3–20.

Kratochwill, T. R., & Brady, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2,* 291–307.

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 December 22, 2010, from What Works Clearinghouse Website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kurtz, P. F., Chin, M. D., Rush, K. S., & Dixon, D. R. (2008). Treatment of challenging behavior exhibited by children with prenatal drug exposure. *Research in Developmental Disabilities, 29,* 582–594. doi:10.1016/j.ridd.2007.05.007

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5,* 87–101.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647–653.

Roscoe, E. M., Carreau, A., MacDonald, J., & Pence, S. T. (2008). Further evaluation of leisure items in the attention condition of functional analyses. *Journal of Applied Behavior Analysis, 41,* 351–364. doi:10.1901/jaba.2008.41-351

Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K. D., 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. doi:10.3758/BRM.41.1.177

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation method: Building the evidence base for method choice in government sponsored evaluation* (pp. 95–109). San Francisco: Jossey-Bass.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 3,* 188–196.

Shadish, W. R., Sullivan, K. J., Hedges, L. V., & Rindskopf, D. M. (2010, November). *A d-estimator for single-case designs*. Paper presented at the conference of A California Consortium for Quantitative Research and Education, Davis, CA.

Stephens, C. E. (2008). Spontaneous imitation by children with autism during a repetitive musical play routine. *Autism, 12*(6), 645–671. doi:10.1177/1362361308097117

Strube, M. J., Gardner, W., & Hartmann, D. P. (1985). Limitations, liabilities and obstacles in reviews of the literature: The current status of meta-analysis. *Journal of Consulting and Clinical Psychology, 5,* 63–68.

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7,* 551–560.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment, 11,* 281–296.