Procedural Sensitivities of Effect Sizes for Single-Case Designs With Directly Observed Behavioral Outcome Measures

James E. Pustejovsky

Funded by:

- U.S. Department of Education, Institute of Educational Sciences

Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures

James E. Pustejovsky

The University of Texas at Austin

Author Note

James E. Pustejovsky, Department of Educational Psychology, University of Texas at Austin, 1912 Speedway, Stop D5800, Austin, TX 78712-1289. Email: pusto@austin.utexas.edu.

This work was supported by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.

An earlier version of this paper was presented at the 2015 annual convention of the American Educational Research Association in Chicago, IL. Supplementary materials for this article are available on the Open Science Framework at https://osf.io/hkzsm/.

Abstract

A wide variety of effect size indices have been proposed for quantifying the magnitude of treatment effects in single-case designs. Commonly used measures include parametric indices such as the standardized mean difference, as well as non-overlap measures such as the percentage of non-overlapping data, improvement rate difference, and non-overlap of all pairs. Currently, little is known about the properties of these indices when applied to behavioral data collected by systematic direct observation, even though systematic direct observation is the most common method for outcome measurement in single-case research. This study uses Monte Carlo simulation to investigate the properties of several widely used single-case effect size measures when applied to systematic direct observation data. Results indicate that the magnitude of the non-overlap measures and of the standardized mean difference can be strongly influenced by procedural details of the study's design, which is a significant limitation to using these indices as effect sizes for meta-analysis of single-case designs. A less widely used parametric index, the log response ratio, has the advantage of being insensitive to sample size and observation session length, although its magnitude is influenced by the use of partial interval recording.

Keywords: effect sizes; meta-analysis; single-case research; non-overlap; behavioral observation; alternating renewal process

Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures

Introduction

Single-case designs (SCDs) are a class of research methods used to evaluate the effects of interventions on individuals. SCDs are defined by the use of repeated measurements of an outcome over time, under distinct treatment conditions or phases, for one or more individual cases. The treatment conditions are deliberately introduced (and, in some designs, removed and re-introduced) by the investigator. Changes in the pattern of outcomes during the phases when the intervention is present compared to when it is absent are taken as evidence that the intervention has a causal effect for that individual. In the logic of single-case research, systematic evidence for an effect accumulates through replication of the pattern at several points in time, across individuals, settings, or target outcomes (Horner et al., 2005).

SCDs comprise a large and important part of the research base in certain areas of psychological and educational research.[1] For example, in a recent, comprehensive review of focused intervention practices for children with autism, 89% of the 456 identified studies used SCDs (Wong et al., 2015). In the field of school psychology, recent systematic reviews on the effects of positive behavioral interventions for reducing challenging behavior among young children and students included predominantly single-case research (e.g., Conroy, Dunlap, Clarke, & Alter, 2005; Maggin, Zurheide, Pickett, & Baillie, 2015). In the field of neuropsychological rehabilitation, Tate, Perdices, McDonald, Togher, and Rosenkoetter (2014) reported that single-case designs comprised over 30% of the research base on

[1]Similar individualized studies with repeated measures are also used in certain areas of medicine and health sciences, where they are often known as *n*-of-1 trials (Duan, Kravitz, & Schmid, 2013; Gabler, Duan, Vohra, & Kravitz, 2011). Although they share common design elements with SCDs, *n*-of-1 trials are used in distinct contexts, are likely to use more sophisticated, high-resolution outcome measurement procedures, and have other distinctive operational features (McDonald et al., 2017). The present investigation therefore limits consideration SCDs as used in psychology and education.

non-pharmacological treatments for acquired brain injury. In light of the size and breadth of the research base in such areas, there is a clear and long-recognized need for principled methods of synthesizing data from SCDs (cf. Allison & Gorman, 1993; Gingerich, 1984).

Traditionally, researchers have drawn conclusions from SCD data based on systematic visual assessment of graphed outcome data (Kratochwill, Levin, Horner, & Swoboda, 2014), and this remains the predominant analytic method for primary SCD studies (Smith, 2012). However, growing interest in evidence-based practice and policy-making has led to renewed attention to methods for statistical analysis and meta-analysis of SCDs, as complements to established visual assessment methods (e.g., Evans, Gast, Perdices, & Manolov, 2014; Maggin & Chafouleas, 2013; Shadish, 2014a; Shadish, Rindskopf, & Hedges, 2008). One focus of recent work has been the development of an array of effect size indices for quantifying the magnitude of intervention effects in SCDs. Effect size indices are the basic unit of analysis in a research synthesis—the metric on which study results are combined and compared. It is therefore imperative for both producers and consumers of research syntheses to understand the interpretation of effect size indices.

Procedural sensitivity

If an effect size index is to have a valid interpretation in terms of intervention effect magnitude, it must provide a reasonable basis for comparison from one study to another (Hedges, 2008; Lipsey & Wilson, 2001). To do so, an effect size should be relatively insensitive to incidental features of how the study was conducted, such as sample size or details of the outcome measurement procedures, which are likely to vary across a collection of studies. An effect size that is instead sensitive to such procedural features can appear to be larger (or smaller) due only to how the study was conducted, rather than because treatment actually produced large (or small) effects. In the context of between-case experiments, procedural insensitivity is one of the primary reasons for using standardized mean differences, rather than t statistics or p values, to summarize and compare results across studies. Whereas t statistics and p values are strongly influenced by sample size,

which is likely to vary from study to study, the standardized mean difference is a function of population parameters only, and thus more comparable across studies (Hedges, 2008). Following similar logic, the psychometric tradition in meta-analysis emphasizes the importance of correcting effect size estimates for "artifacts," such as direct or indirect range restrictions and differential reliability, in order to reduce procedural sensitivity and improve comparability across studies that use varied procedures (Hunter & Schmidt, 2004).

Using a procedurally sensitive effect size for purposes of meta-analysis has at least two important consequences. First, it harms the basic intepretability of the synthesis because the metric on which results are combined or compared is not uniform. Second, to the extent that a collection of studies includes a variety of operational procedures, using a procedurally sensitive effect size adds extraneous variation, making it more difficult to identify substantively meaningful moderating factors. Procedural insensitivity is thus a basic and fundamental property of an effect size index, required if it is to be useful for synthesizing results across multiple studies. However, insufficient attention has been paid to whether effect sizes that are commonly used for summarizing the results of SCDs have this property.

SCD effect sizes

A wide array of effect sizes have been proposed for use with SCDs, yet there remains considerable disagreement regarding their merits. Some of the effect sizes, including within-case standardized mean differences (Busk & Serlin, 1992; Gingerich, 1984), between-case standardized mean differences (Shadish, Hedges, & Pustejovsky, 2014), and measures based on piece-wise linear regression models (e.g., Center, Skiba, & Casey, 1985-86; Maggin, Swaminathan, et al., 2011), are based on parametric statistical models. Most of these models are premised on the assumption that the dependent variable is normally distributed, yet this assumption is often criticized as inappropriate for data from SCDs (cf. Solomon, Howard, & Stein, 2015). Little previous research has examined the performance of these effect sizes when the assumed statistical model is not correct.

The other main family of effect sizes for single-case research are the non-overlap measures (NOMs), which include the percentage of non-overlapping data (Scruggs, Mastropieri, & Casto, 1987), improvement rate difference (Parker, Vannest, & Brown, 2009), and a number of others. One perceived advantage of these measures is that they are non-parametric, in the sense that they are not based upon distributional assumptions about the dependent variable (Parker, Vannest, & Davis, 2011). NOMs are also viewed as being intuitively interpretable because they are defined in terms of overlap percentages and are on a scale of 0 to 100%. In part because of these perceived advantages, the NOMs are the most widely employed effect sizes in reviews and syntheses of single-case research (Maggin, O'Keeffe, & Johnson, 2011). However, the fact that the NOMs are not developed under specific assumptions about the distribution of the data makes it more difficult to determine whether—and under what circumstances—these indices are sensitive to procedural aspects of a study's design. The aim of the present study is to fill this gap by studying the characteristics of the NOMs and of parametric effect size indices using a realistic model for direct observation of behavior.

Behavioral observation data

Behavioral measures derived from systematic direct observation are the most common type of dependent variable in single-case research (Gast, 2010). Systematic direct observation entails watching the behavior of a participant or participants for a specified period of time (an observation session) and using a recording system to score the occurrence, duration, or other features of a behavior. Single-case research emphasizes the importance of careful operational definition of the focal behavior and collection of inter-observer reliability data to ensure the validity of the resulting measurements (Horner et al., 2005). A variety of different systems are used to record direct observation of behavior, including continuous recording, momentary time sampling, frequency counting, and partial interval recording (Ayres & Gast, 2010). Each of these procedures can be used for shorter or longer observation sessions, and the number of observation sessions will also

vary from study to study. Thus, if an effect size index is to provide a fair basis for comparing SCDs that use behavioral outcomes, it should be relatively insensitive to the researcher's decisions about how the dependent variable is measured, as well as to decisions about other aspects of the study's design.

In order to study the properties of effect size indices when applied to behavioral outcome measures, a means of simulating realistic behavioral observation data is needed. A useful tool for doing so is the alternating renewal process, which is a statistical model for the stream of behavior as it is perceived during an observation session (Pustejovsky & Runyon, 2014; Rogosa & Ghandour, 1991). A key benefit of using this model is that it mimics the physical process of observing a behavior stream and recording data as one does so. Consequently, it provides a way to emulate several distinctive features of real, empirical behavioral observation data.

Aim and scope

Using the alternating renewal process model, this study investigates the extent to which several proposed effect sizes for SCDs, including six NOMs and two parametric indices, are sensitive to procedural features of single-case studies. The procedural features investigated include the number of observation sessions in the baseline and treatment phases, the length of the observation sessions, and the recording system used to collect measurements of behavior. The study focuses on these procedural features because they all reflect basic operational decisions that must be made when designing a single-case study and because they are likely to vary across a collection of SCDs on a common topic—or even across cases within a single study. Other procedural features, such as the operational definition of the focal behavior and degree of inter-observer reliability, might also be of concern when synthesizing a collection of SCDs, but remain outside the scope of the present investigation. I consider the implications of this scope limitation in the discussion section.

The scope of the study is also limited in two ways with respect to the effect size indices under consideration. First, in order to isolate the basic interpretation and

procedural sensitivity of the effect sizes, the review and subsequent simulation are limited to effect sizes that are appropriate for data without systematic time trends. Accounting for time trends is important in many applications. However, the extant effect size indices that do account for time trends are all extensions of the basic, widely used effect sizes included in this review (e.g., piece-wise linear regression models extend the within-case standardized mean difference). Consequently, they are very likely to retain the same interpretation—and have similar procedural sensitivities—as the indices upon which they are built.[2]

Second, consistent with the idiographic orientation of single-case research, the review focuses only on effect sizes that quantify treatment effects for individual cases, considered separately. Other recently proposed effect sizes such as between-case standardized mean differences (Shadish et al., 2014) are measures of *average* effects across multiple individual cases, designed to achieve comparability with average effect sizes from between-group designs. Thus, they serve a distinct purpose from case-specific indices. Investigating the procedural sensitivities of these more complex metrics would require a more elaborate simulation model that describes both between- and within-case variation; it therefore remains a topic for future research.

Given these scope limitations, the review includes the following NOMs: the percentage of non-overlapping data, the percentage exceeding the median, the percentage of all non-overlapping data, the robust improvement rate difference, the non-overlap of all pairs, and the Tau index. Parker, Vannest, and Davis (2011) provide a more expansive review of the NOMs, including worked examples of how to calculate each effect size based on graphed data. The review also includes two parametric indices: the within-case standardized mean difference and the log response ratio, an established effect size for between-case designs that has recently been proposed for use with SCDs. Pustejovsky and Ferron (2016) provide a more detailed discussion of both parametric measures. Taken together, the present review includes the most well-known and widely used effect size

[2]Section S2 of the supplementary materials examines this point in greater detail.

indices in single-case research, so that its findings are relevant to current practices for conducting systematic reviews of single-case research.

The remainder of the manuscript is organized as follows. The next section reviews several effect size indices that have been proposed for use with SCDs. The following three sections describe, respectively, the alternating renewal process model used to simulate behavior data, the design of the simulation study, and the simulation results. The final section discusses limitations, implications for synthesis of single-case research, and avenues for future research.

Calculating effect sizes for SCDs

This section describes six NOMs and two parametric indices that have been proposed for use with SCDs. In addition to defining each measure, I note the range of possible values and the null value for each index (i.e., the expected value when treatment has no effect), as well as any available guidelines for characterizing effects as "small," "medium," or "large." Such benchmarks provide one way to judge whether the procedural sensitivities of the effect sizes are consequential. Table 1 summarizes the properties of the indices under consideration. Section S1 of the supplementary materials includes numerical examples for all of the effect size indices described in this section.

Notation

Each of the effect size indices is defined in terms of a comparison between a single baseline phase and a single treatment phase. Let m denote the number of observations in the baseline phase and n denote the number of observations in the treatment phase. Denote the outcome measurements during the baseline phase as $y_1^A, ..., y_m^A$ and the outcome measurements during the treatment phase as $y_1^B, ..., y_n^B$. For the NOMs, the following definitions assume that the dependent variable is operationalized such that smaller values correspond to more beneficial outcomes, so that decreases are desirable.[3] Let $I(E)$ denote

[3]For outcomes where an increase is desirable, one would first multiply the outcome by -1 and then evaluate the specified formula for the NOM.

Table 1

Range, null value, and benchmarks for SCD effect size indices

Index	Minimum	Maximum	Null value	Benchmarks*
PND	0%	100%	$\frac{100\%}{m+1}$	50% / 70% / 90%
PEM	0%	100%	50%	none
PAND	$100\% \times \frac{\max\{m,n\}}{m+n}$	100%	dependent on m, n	none
RIRD	$\frac{1}{2} - \min\left\{\frac{m}{2n}, \frac{n}{2m}\right\}$	1	dependent on m, n	.5 / .7
NAP	0%	100%	50%	65% / 92%
Tau	-1	1	0	.3 / .84
SMD	$-\infty$	∞	0	1.0 / 2.5
LRR	$-\infty$	∞	0	none

* The Benchmarks column reports the cut-off values between different categorical labels for characterizing the magnitude of an effect size index. PND = percentage of non-overlapping data; PEM = percentage exceeding the median; PAND = percentage of all non-overlapping data; RIRD = robust improvement rate difference; NAP = non-overlap of all pairs; SMD = standardized mean difference; LRR = log response ratio.

the indicator function, which is equal to one when condition E is true and equal to zero when E is false.

Non-overlap measures

Percentage of non-overlapping data. The percentage of non-overlapping data (PND) was the first non-overlap measure to appear in the literature. It is defined as the percentage of measurements in the treatment phase that are less than the lowest measurement from the baseline phase (Scruggs et al., 1987). Mathematically,

$$\text{PND} = 100\% \times \frac{1}{n}\sum_{i=1}^{n} I\left(y_i^B < y_{(1)}^A\right), \tag{1}$$

where $y_{(1)}^A = \min\left\{y_1^A, ..., y_m^A\right\}$. PND can take on values between 0 and 100%. Scruggs and Mastropieri (1998) offered general guidelines for the interpretation of PND, suggesting that a PND value of 90% or greater could be interpreted as indicating a "very effective"

intervention; a PND between 70% and 90% as indicating an "effective" one; a PND between 50% and 70% as indicating a "questionable" effect; and a PND of less than 50% as indicating an "ineffective" intervention (p. 224). Manolov and Solanas (2009) proposed an extension of PND that accounts for baseline time trends.

Since it was first proposed, PND has been widely criticized (e.g., Shadish et al., 2008; O. R. White, 1987; Wolery, Busick, Reichow, & Barton, 2010). In an analysis similar to the simulations presented in a later section, Allison and Gorman (1994) demonstrated that the expected value of the PND statistic is strongly influenced by the number of observations in the baseline phase, with longer baseline phases tending to result in smaller values of PND, even when treatment has no effect at all.[4] They argued that this dependence on sample size makes the statistic unsuitable for use as an effect size metric. Despite this and other objections, PND remains by far the most commonly applied effect size in systematic reviews of SCDs (Maggin, O'Keeffe, & Johnson, 2011; Scruggs & Mastropieri, 2013).

Percentage exceeding the median. To address some of the criticisms of PND, Ma (2006) proposed an alternative that uses the median of the baseline phase (rather than the minimum) as the basis for comparison with the treatment phase. For an outcome where increase is desirable, the percentage exceeding the median (PEM) is defined as the percentage of measurements in the treatment phase that exceed the median of the baseline phase measurements. For an outcome where decrease is desirable, PEM is defined as the percentage of treatment phase measurements that are less than the median of the baseline phase. To account for the possibility of ties in the data, measurements in the treatment phase that are exactly equal to the median of the baseline phase are counted as half an

[4] Allison and Gorman (1994) indicated that the expected null value of PND is $100\% \times \left(1 - 2^{-1/n}\right)$. However, this is incorrect. When the outcome measure is continuous, the expected null value of PND is $100\%/(m+1)$. For discrete outcome measures, the null value of PND can be slightly different due to the possibility of ties.

observation. For an outcome where decrease is desirable, PEM is calculated as

$$\text{PEM} = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\left[I\left(y_i^B < \tilde{y}_A\right) + 0.5I\left(y_i^B = \tilde{y}_A\right)\right], \tag{2}$$

where $\tilde{y}_A = \text{median}\left\{y_1^A, ..., y_m^A\right\}$. Like PND, PEM ranges in principle from 0 to 100%. Unlike PND, the expected value of PEM is stable when treatment has no effect: if the outcomes in the treatment phase are distributed just as the outcomes in the baseline phase, then the expected value of PEM will be 50%. To my knowledge, no guidance has been offered regarding what constitutes a small, medium, or large value of PEM. For handling baseline time trends, Wolery et al. (2010) proposed calculating the percentage of treatment phase observations that exceed a split-middle trend line.

Percentage of all non-overlapping data. Parker, Hagan-Burke, and Vannest (2007) proposed the percentage of all non-overlapping data (PAND) as another alternative to PND. As originally described, the percentage of all non-overlapping data (PAND) is defined as 100% minus the minimum percentage of observations that would need to be swapped between the baseline and treatment phases so that the lowest measurement in the baseline phase exceeds the highest measurement in the treatment phase. More recent descriptions use a subtly different definition, defining PAND as the percentage of observations remaining after removing (instead of swapping) the fewest possible number of observations from either phase so that the lowest remaining point from the baseline phase exceeds the highest remaining point from the treatment phase (Parker, Vannest, & Davis, 2011, 2014). The simulation study employs the latter definition on the assumption that it supersedes the former.

Let $y_{(1)}^A, y_{(2)}^A, ..., y_{(m)}^A$ denote the values of the baseline phase data, sorted in increasing order, and let $y_{(1)}^B, y_{(2)}^B, ..., y_{(n)}^B$ denote the values of the sorted treatment phase data. For notational convenience, let $y_{(m+1)}^A = \infty$ and $y_{(0)}^B = -\infty$. Let

$$x = \max\left\{(i+j)\, I\left(y_{(m+1-i)}^A > y_{(j)}^B\right)\right\}, \tag{3}$$

where the maximum is taken over the values $0 \leq i \leq m$ and $0 \leq j \leq n$. Then

$$\text{PAND} = 100\% \times x/(m+n). \tag{4}$$

Unlike PND and PEM, the range of the PAND statistic is not obvious. If there is complete separation between phases, then PAND will equal 100%. However, the minimum possible value of PAND is not 0 (as might be expected), but rather $100\% \times \max\{m, n\}/(m+n)$, the number of observations in the longer of the two phases, divided by the total number of observations.[5] Although Parker et al. (2007) indicated that 50% is the expected value of PAND when the intervention has no effect on the outcome, this cannot be the case because, when m is not equal to n, the minimum possible value is larger than 50%. Parker and colleagues (Parker et al., 2007; Parker, Vannest, & Davis, 2011) reported the empirical distribution of PAND effect sizes based on large samples of published SCDs, but did not offer any interpretation guidelines.

Robust improvement rate difference. Parker et al. (2009) described the "robust improvement rate difference" (RIRD), which is equivalent to the robust phi coefficient corresponding to a 2×2 table arrangement of the numbers obtained in calculating PAND (Parker, Vannest, & Davis, 2011). With x defined as in Equation (3), RIRD is calculated as

$$\text{RIRD} = \frac{n - x/2}{n} - \frac{x/2}{m}. \tag{5}$$

RIRD is a linear re-scaling of PAND, such that

$$\text{RIRD} = \frac{1}{2mn}\left[(m+n)^2 \frac{\text{PAND}}{100\%} - m^2 - n^2\right].$$

The range of RIRD therefore depends on the ratio of m to n. As with PAND, the expected value of RIRD when the intervention has no effect on the outcome is unclear; the

[5] Consider the case where the maximum of the baseline phase is less than the minimum of the treatment phase. To obtain no overlap, one must either remove all baseline observations or all treatment observations; thus, the minimum number of observations that must be removed is equal to the number of observations in the shorter of the two phases, and the number of observations remaining is equal to the number of observations in the longer phase.

simulation results presented in a later section indicate that the null value actually depends on the number of observations in each phase. Based on a comparison between RIRD values and expert visual assessments, Parker et al. (2009, p. 147) provided tentative benchmarks for RIRD, suggesting that values below .50 correspond to "questionable" effects, values between .50 and .70 correspond to "medium" effects, and values above .70 correspond to "large" effects.

Non-overlap of all pairs. Parker and Vannest (2009) proposed the non-overlap of all pairs (NAP) statistic, which involves pairwise comparisons between each point in the treatment phase and each point in the baseline phase. NAP is defined as the percentage of all such pairwise comparisons where the measurement from the treatment phase is less than the measurement from the baseline phase, with pairs of data points that are exactly tied being given a weight of 0.5. Mathematically,

$$\text{NAP} = 100\% \times \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[I\left(y_j^B < y_i^A\right) + 0.5 I\left(y_j^B = y_i^A\right) \right]. \tag{6}$$

NAP can take on values between 0 and 100% and has a stable null value of 50%.

Parker and Vannest (2009) argued that NAP has several advantages over other non-overlap measures, including ease of calculation, better discrimination among effects in published SCDs, and the availability of valid confidence intervals. As they also noted, NAP has been proposed as an effect size index (under a variety of different names) in many other areas of application (e.g., Vargha & Delaney, 2000). Based on visual assessment of a corpus of SCD studies, Parker and Vannest (2009) characterized NAP values of less than 65% as "weak," values between 66% and 92% as "medium," and values between 93% and 100% as "large" (p. 364). Other empirical studies have suggested alternative cut-offs (Petersen-Brown, Karich, & Symons, 2012; Solomon et al., 2015).

Tau. Parker, Vannest, Davis, and Sauber (2011) described the Tau effect size, which is closely related to NAP but can be extended to handle time trends in the baseline and treatment phases. Without adjustment for time trends, Tau is simply a linear

re-scaling of NAP to have a range of -1 to 1:

$$\text{Tau} = 2 \times \frac{\text{NAP}}{100\%} - 1. \tag{7}$$

Tau has an expected value of 0 when treatment has no effect on the outcome. Benchmark values for Tau, as listed in Table 1, can be derived from the benchmarks for NAP proposed by Parker and Vannest (2009) based on the algebraic relationship between the indices. Because NAP and Tau are so closely related, the simulation study focuses on the former measure only. Parker, Vannest, Davis, and Sauber (2011) and Tarlow (2017) proposed different extensions to Tau that account for time trends.

Parametric measures

Standardized mean difference. The standardized mean difference (SMD) is a widely used effect size measure for synthesis of between-case experimental designs. In a study of two independent groups, the SMD parameter is defined as the difference in population means, scaled by the standard deviation of the outcome, which is typically assumed to be constant across groups (Borenstein, 2009). Gingerich (1984) proposed to use a version of the standardized mean difference, defined in terms of within-case variation, for quantifying effect sizes in SCDs. Let $\bar{y}_A$ and $\bar{y}_B$ denote the sample means of the baseline and treatment phases, respectively, and let s_A^2 and s_B^2 denote the sample variances of the baseline and treatment phases. Gingerich (1984) proposed to use

$$d = \left(\bar{y}_B - \bar{y}_A\right) / s_A \tag{8}$$

as an estimator of the within-case SMD (see also Busk & Serlin, 1992).

Maggin, O'Keeffe, and Johnson (2011) reported that the within-case SMD was the second most frequently employed effect size in meta-analyses of single-case research on students with disabilities. Harrington and Velicer (2015) applied the within-case SMD to a corpus of single-case studies published in the *Journal of Applied Behavior Analysis* in 2010. Based on the distribution of observed effect size estimates, they proposed that d statistics

between 0 and 1 could be characterized as "small," between 1 and 2.5 as "medium," and greater than 2.5 as "large." Maggin, Swaminathan, et al. (2011) described an extension of the within-case SMD that uses a piece-wise linear regression model to account for time trends.

The within-case d estimator has a small-sample bias that makes its magnitude sensitive to the length of the baseline phase. The simulation study reported in a later section therefore examines a bias-corrected SMD estimator, calculated as

$$g = \left(1 - \frac{3}{4n_A - 5}\right)\left(\frac{\bar{y}_B - \bar{y}_A}{s_A}\right) \tag{9}$$

(cf. Hedges, 1981). This multiplicative correction factor is expected to remove the small-sample bias of d, thereby reducing its procedural sensitivity.

Log response ratio. The log response ratio (LRR) effect size index quantifies the magnitude of treatment effects in terms of proportionate change in the level of an outcome. Its use is well-established in certain domains of between-groups research (e.g., Hedges, Gurevitch, & Curtis, 1999). Pustejovsky (2015) proposed the LRR as an effect size measure for SCDs with behavioral outcome measures. Letting μ_A and μ_B be the average levels of the outcome in the baseline and treatment phases, respectively, the LRR parameter is defined as $\lambda = \ln(\mu_B/\mu_A)$, where ln() denotes the natural logarithm function. Because the LRR parameter is a function of the means within each phase, its magnitude captures changes in levels only (i.e., basic effects), rather than other features of the data that might be considered in visual analysis, such as changes in variability.

Because the LRR quantifies change in proportionate terms, it can be interpreted by converting it into a percentage change in the outcome from baseline to intervention, using the formula: % Change $= 100\% \times \left(e^{\lambda} - 1\right)$ (Pustejovsky, 2015). Campbell and Herzinger (2010) argued that proportionate change measures of effect size are intuitively appealing because applied researchers and clinicians commonly conceptualize and discuss treatment impacts in such terms. Furthermore, percentage change measures that are conceptually similar to the LRR have occasionally been used in syntheses SCDs, under the names of the

suppression index (Marquis et al., 2000) or the mean baseline reduction (Campbell, 2003).

For outcomes that are quantified as proportions (or percentages), the magnitude of the LRR depends on the direction of therapeutic improvement. When calculating the LRR for a set of several studies, the analyst must therefore ensure that outcome measures are all defined using a consistent direction of therapeutic improvement, so as to avoid introducing procedural sensitivity.[6] With outcomes that are defined in a consistent direction, a basic moment estimator of the LRR can be calculated as

$$R_1 = \ln\left(\bar{y}_B\right) - \ln\left(\bar{y}_A\right). \tag{10}$$

Like the within-case d index, this basic moment estimator has a small-sample bias that will make its magnitude sensitive to the number of observations in each phase. A bias-corrected estimator of the LRR is given by

$$R_2 = \ln\left(\bar{y}_B\right) + \frac{s_B^2}{2n_B\bar{y}_B^2} - \ln\left(\bar{y}^A\right) - \frac{s_A^2}{2n_A\bar{y}_A^2}. \tag{11}$$

(Pustejovsky, 2015).[7] The simulation study examines the performance of both estimators.

The alternating renewal process model

To study the properties of the NOM and parametric effect sizes when applied to behavioral observation data, a statistical model is needed that is flexible enough to capture the relationships between the procedures used to measure a behavior (i.e., observation session length, choice of recording system) and the distribution of the resulting data. Common probability models such as the normal, binomial, or Poisson distribution are not well-suited for modeling these relationships. The simulations described in the following

[6] Pustejovsky (2018) provides more detailed explanations and instructions for calculating the LRR and estimating its sampling variance when conducting a meta-analysis of SCDs.

[7] In practice, the sample means must be truncated at a small value so that R_1 and R_2 remain defined. In the simulation study, I truncated the sample mean for phase $p = A, B$ at the value $1/(2kn_p)$, where k is the total number of intervals per session. For continuous recording, k is set to the length of the observation session in seconds.

section therefore used a more complex—yet also more realistic and flexible—model called the alternating renewal process.

Before explaining the details of the model, it is useful to distinguish between two classes of behavior: state behaviors and event behaviors. State (or duration-based) behaviors consist of episodes that each last some length of time. With such behaviors, the researcher is typically most interested in the behavior's prevalence, meaning the overall proportion of time that it occurs. A continuous recording or interval recording system is often used for measurement of state behaviors. Percentage of time on-task is a common example of a state behavior. In contrast, event (or frequency-based) behaviors consist of behavioral events that each have negligible duration. With such behaviors, the researcher is typically most interested in the behavior's incidence, meaning the overall rate of occurrence per unit of time, and frequency counting is often used for measurement. Hitting and biting are common examples of event behaviors. Both state behaviors and event behaviors can be simulated using the alternating renewal process model.

The alternating renewal process model simulates behavioral observation data using a two-step process. The first step is to simulate a behavior stream for each measurement occasion. Each behavior stream consists of episodes of behavior separated by spans of inter-response time, emulating the pattern of behavior that an observer would actually see during the course of an observation session. The duration of each behavioral episode is generated randomly, according to a certain probability distribution (e.g., an exponential distribution); similarly, the length of each inter-response time is generated randomly, according to a different probability distribution. The process of alternately generating behavioral episode durations and inter-response times is repeated until their cumulative sum meets or exceeds the length of the observation session.

The second step is to calculate a summary measurement for each of the simulated behavior streams, based on the rules of a specific recording system. This process mimics the steps that an observer follows as they record data during a session and then calculate a

summary of that data. Each recording system is modeled using a different set of rules. For frequency counting, the summary measurement is the total number of episodes of behavior during the session. For continuous recording, the summary measurement is percentage duration, calculated as 100% times the sum of all the episode durations, divided by the total session length. For interval recording systems such as momentary time sampling (MTS) or partial interval recording (PIR), the observation session is divided up into short intervals of time, each of a specified length (e.g., 15 seconds). Each interval is scored according to a set of rules for determining whether the behavior is present. In MTS, the behavior is scored as present if is occurring at the very end of the interval; in PIR, the behavior is scored as present if it occurs at any point during the interval. With both interval systems, a summary measurement is calculated as the percentage of intervals where the behavior was present. Pustejovsky and Runyon (2014) provide further details about the recording systems and summary measurements, as well as examples of simulated behavior streams.

Simulation design

Using the alternating renewal process model, I conducted a simulation study to examine the extent to which SCD effect size indices are sensitive to procedural features of behavioral observation data, which are likely to vary across a collection of SCDs to be synthesized. The simulations reported in the following section focus on state behaviors, where the behavior's prevalence is the primary characteristic of interest. Another simulation study focusing on event behaviors is reported in the supplementary materials.[8]

Both simulation studies involved generating data from a single pair of phases within an SCD, where the expected value of the outcome (i.e., the average level of behavior) was stationary within each phase but could change between phases. This pair of phases might represent the baseline and treatment phases for a case within a multiple baseline design, or

[8]The simulation of state behaviors is presented in the main text because it is larger, involving seven different recording systems, whereas the event behavior simulation involves only four recording systems.

different conditions in an alternating treatment design. It might also represent one pair of phases within a treatment reversal (ABAB) design that includes several further phase contrasts. Although designs that include multiple pairs of phases are important in practice, the simulations focus on this simplified scenario for two reasons. First, all of the existing effect size indices are defined in terms of a single phase pair, and so following the same structure in the simulations provides the clearest way to evaluate the effect sizes as originally defined. Second, calculating effect sizes from SCDs multiple phase contrasts involves averaging effect sizes calculated from single pairs of phases, and so effect sizes calculated from more complicated designs will share the same procedural sensitivities as those calculated from the simpler design considered here.[9] Similarly, the simulation used a data-generating model without systematic time trends because this is a common assumption of all the indices under consideration. In summary, simulating single phase-pairs without time trends provided the simplest and most direct means of evaluating the procedural sensitivities of the indices as defined, under the conditions for which they were designed.

Data-generating model

The simulation used the alternating renewal process model to generate realistic measurements of a state behavior, in the context of a single phase-pair within an SCD study. Several further details had to be specified in order to fully operationalize the alternating renewal process model. Although prevalence is the main characteristic of a state behavior, its incidence is also relevant because incidence influences the variability of measurements of the behavior. I assumed that the behavior's prevalence and incidence were constant within each phase of the study, but could change between phases. The prevalence and incidence of the behavior within a given phase determined the average

[9]Section S3 of the supplementary materials reviews typical approaches for calculating effect sizes from designs with multiple phase contrasts and provides a more detailed explanation of why they will share the same procedural sensitivities.

episode duration and average inter-response time used to generate behavior streams. I assumed that episode durations and inter-response times followed exponential distributions. I selected exponential distributions purely for sake of simplicity, as they are one-parameter, continuous distributions on the positive real line.[10]

I chose values for the behavioral parameters to represent a range of realistic conditions. I set the prevalence of the behavior during the baseline phase to 20%, 50%, or 80% in order to capture a range of different types of behavior, such as mild, moderate, or severe problem behavior. For incidence, Mudford, Locke, and Jeffrey (2011) reported that SCDs published in the *Journal of Applied Behavior Analysis* between 1998 and 2007 displayed a median rate of responding of slightly less than once per minute, with a maximum rate well above once per minute in almost all cases. Based on their findings, I set the baseline incidence of the behavior to once per minute or twice per minute. Many SCDs focus on behaviors for which a decrease is desirable; I therefore simulated data in which the treatment reduces the prevalence and incidence of the behavior by 0% (representing no effect of treatment), 50%, or 80% (representing a substantial decrease in the behavior). Figure S2 in the supplementary materials displays examples of simulated SCDs for each combination of prevalence, and incidence, and change in behavior.

Procedural factors

The simulation examined the effects of three procedural factors: recording system, length of observation session, and number of observations in the baseline and treatment phases. In order to test the effect size indices under realistic conditions, I selected levels for these factors that closely resemble the procedures used in actual SCDs.

Continuous recording (CR), MTS, and PIR are the main systems for recording direct

[10]The event behavior simulations investigated a wider set of distributions, including both exponential and gamma distributions. For the state behavior simulations, the theory of alternating renewal processes (Rogosa & Ghandour, 1991) would suggest that using other distributions, such as the two-parameter gamma distribution, may influence the degree of procedural sensitivity with respect to different partial interval recording procedures, but would not strongly affect sensitivity to other procedural factors.

observation of a state behavior (Ayres & Gast, 2010). Reviews of the single-case literature indicate that all three of these procedures are used in practice (Adamson & Wachsmuth, 2014; Mudford, Taylor, & Martin, 2009). For the interval-based systems, commonly used interval lengths are 10, 15, 20, or 30 s. The simulation therefore examined CR; MTS with 10, 20, or 30 s intervals; and PIR with 10, 20, or 30 s intervals.

Any of these recording systems may be used for longer or shorter observation sessions. For example, in a large synthesis of SCDs examining the effect of functional behavior assessment interventions on student problem behavior (Gage, Lewis, & Stichter, 2012), observation session lengths ranged from 5 to 60 min and 75% of cases were observed for 20 min or less. To emulate conditions typically used in practice, the simulation examined observations sessions lasting 5, 10, 15, or 20 min. Finally, the number of observations per phase ranges widely in SCDs, with some phases consisting of fewer than 5 measurement occasions while others including far more. In a review of over 400 SCDs published between 2000 and 2010, Smith (2012) found that baseline phases included an average of 10.2 observations, with a range of 1 to 89. In a review of 112 SCDs published in 2008, Shadish and Sullivan (2011) reported that the majority of cases used initial baselines of 5 or more observations. The simulations therefore examined designs with 5, 10, 15, or 20 sessions in the baseline phase and 5, 10, 15, or 20 sessions in the treatment phase, including all 16 possible combinations across the two phases.

Simulation procedures and analysis

I conducted the simulation using the ARPobservation package (Pustejovsky, 2016) for the R statistical computing environment. The computer code that implements the simulation and full numerical results are available in the supplementary materials. The simulation used a full factorial, $4 \times 4 \times 4 \times 3 \times 2 \times 7$ design. Table 2 summarizes the parameters and levels of the simulation design. For each combination of factor levels, I simulated 2000 phase-pairs and calculated each of the effect size measures.

To analyze the simulation results, an operational definition of procedural sensitivity

Table 2

State behavior simulation design

Parameter	Levels
Prevalence	0.2, 0.5, 0.8
Incidence (per min)	1, 2
Change (% decrease)	0%, 50%, 80%
Session length (min)	5, 10, 15, 20
Sessions in the baseline phase	5, 10, 15, 20
Sessions in the treatment phase	5, 10, 15, 20
Recording system	CR, MTS (10, 20, 30 s), PIR (10, 20, 30 s)

CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

is needed. I operationalized procedural sensitivity as the *conditional range of the expected value* of each effect size, where the range is taken across the levels of a given procedural factor while holding all other factors constant. To understand this summary measure, it is helpful to consider a somewhat simpler experiment in which only three factors are manipulated. Let μ_{fgh} denote the expected value of a given effect size at a given combination of factor levels $f = 1, ..., F$; $g = 1, ..., G$; and $h = 1, ..., H$ (for instance, f might be the number of observations in the baseline phase, g the length of the observation session, and h the combination of all other factors in the simulation). I first estimated μ_{fgh} by taking the average of each effect size index across replications.[11] I then calculated the conditional range of the effect size index across a procedural factor f, holding constant the other factors g and h, by taking the difference between the maximum and minimum

[11] Using 2000 replications yielded precise estimates of the expected values, with negligible Monte Carlo error.

expected value across the levels of f:

$$\max\left\{\mu_{1gh}, \mu_{2gh}..., \mu_{Fgh}\right\} - \min\left\{\mu_{1gh}, \mu_{2gh}..., \mu_{Fgh}\right\},$$

for each $g = 1, ..., G$ and $h = 1, ..., H$. If the expected value of an effect size is entirely unaffected by factor f—that is, if it is *insensitive* to f—then its conditional range will be zero. Conversely, the more strongly that the magnitude of an effect size is affected by factor f, the larger will be its conditional range. To the extent that there are interactions among effects, the conditional range with respect to a given factor will vary depending on the levels of the other factors. To summarize the overall sensitivity of an effect size metric to a procedural factor f, I created violin plots representing the full distribution of conditional range values across the levels of g and h.

In addition to summary plots, I examined the expected values of each effect size directly for subsets of the simulation conditions as a means of understanding the exact range of conditions under which an effect size is procedurally sensitive. These results are presented in the form of figures that illustrate how the expected value of an effect size varies as a function of the procedural factors in the design. For effect sizes whose expected values are not influenced by the number of sessions in the baseline or treatment phase, results are averaged across the levels of one or both of these factors. For clarity of presentation, some of the figures display results for selected subsets of the conditions; in these cases, the results presented are generally consistent with the other simulation conditions except when otherwise noted.

In interpreting the simulation results, it is helpful to have guidelines for assessing the extent to which an effect size is sensitive to a given procedural factor. For the NOMs, which have finite ranges, one means of making such judgments is to compare the extent of sensitivity to the range of possible values for the index. I characterized an effect size as "sensitive" to a given procedural factor if that factor influenced its magnitude by at least 10% of this range. For all of the NOMs except RIRD, 10% of the range is equivalent to 10 percentage points or less; for RIRD, 10% of the range is equivalent to .15 or less. This

approach is roughly consistent with the interpretive benchmarks that have been proposed for some of the NOMs, in that a procedural factor that can influence the effect size by 10% or more creates the possibility of shifting an interpretation from "weak" to "medium" or from "medium" to "strong." This approach is not applicable to the parametric effect sizes, whose ranges are not bounded. Instead, I classified SMD effect sizes as "sensitive" to a procedural factor that could influence its magnitude by 0.5, which is roughly consistent with the interpretive guidelines proposed by Harrington and Velicer (2015). I classified LRR effect sizes as "sensitive" to a procedural factor that produced changes in magnitude of 0.10 or more.

Simulation results

Non-overlap measures

Figure 1 depicts the conditional range distributions for each of the NOMs with respect to the four procedural factors in the simulation: number of sessions in the baseline phase, number of sessions in the treatment phase, observation session length, and recording system. Each column corresponds to a procedural factor; each row to one of the effect size measures. Within each panel, the full distribution of conditional ranges is represented using a violin plot, where width corresponds to relative frequency of a given value for the conditional range; the horizontal bars within the violin plot correspond to the quintiles of the distribution. Violin plots with substantial density at high values of the conditional range indicate that an effect size measure is sensitive to a procedural factor under a range of conditions. The conditional range distributions are plotted separately by the actual percentage change in behavior in order to assess procedural sensitivity for null versus beneficial effects. I describe results for each of the NOMs in turn.

PND. Results in the top row of Figure 1 indicate that PND is sensitive to the number of sessions in the baseline phase, observation session length, and recording system, but not to the number of sessions in the treatment phase. Results for a change of 0% (in the first column of each panel) are consistent with the findings of Allison and Gorman

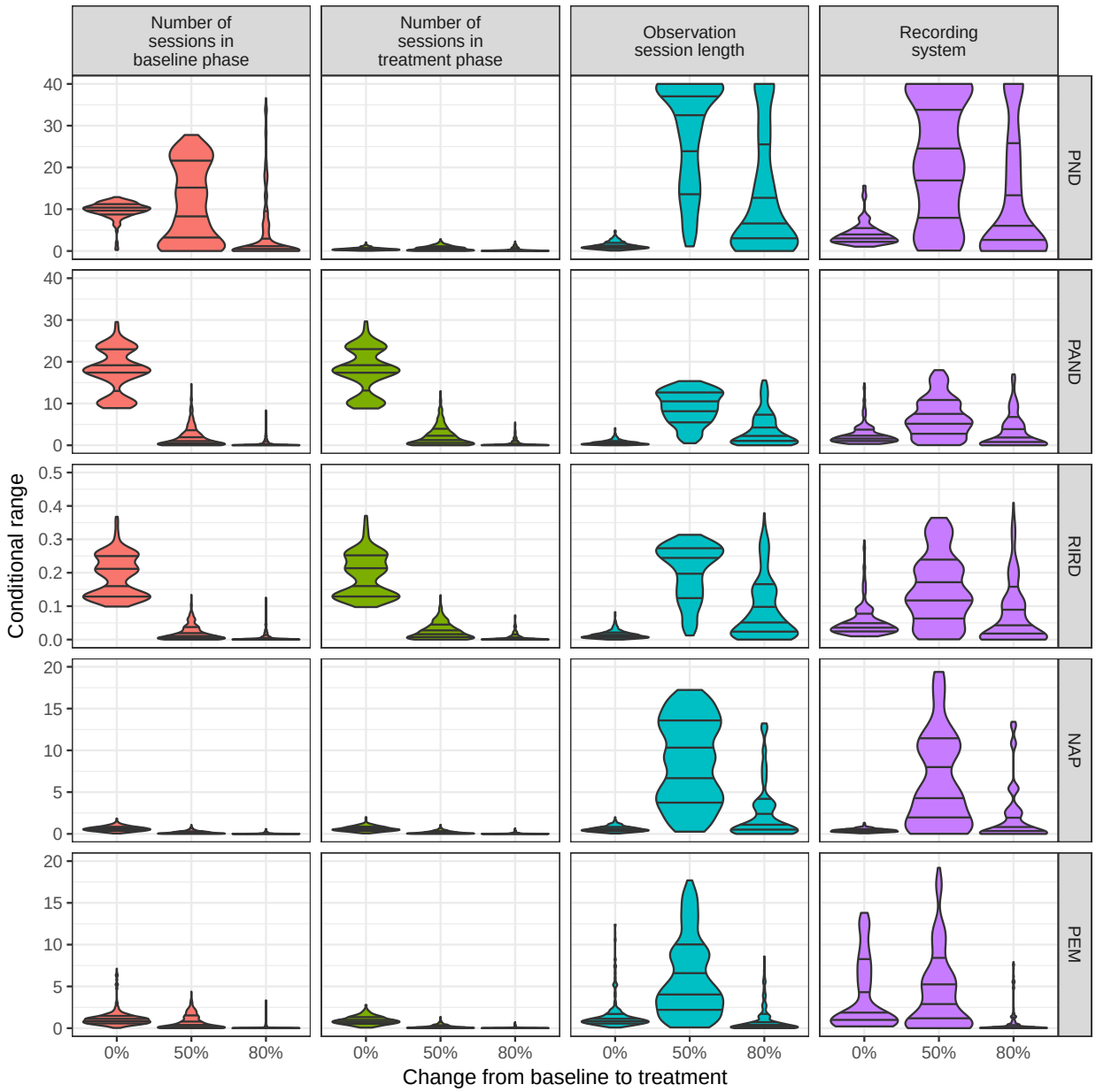

Figure 1. Conditional range distributions of the non-overlap effect size measures for each procedural factor, by percentage change from baseline to treatment. For clarity of illustration, the conditional range distributions for PND are truncated at 40.

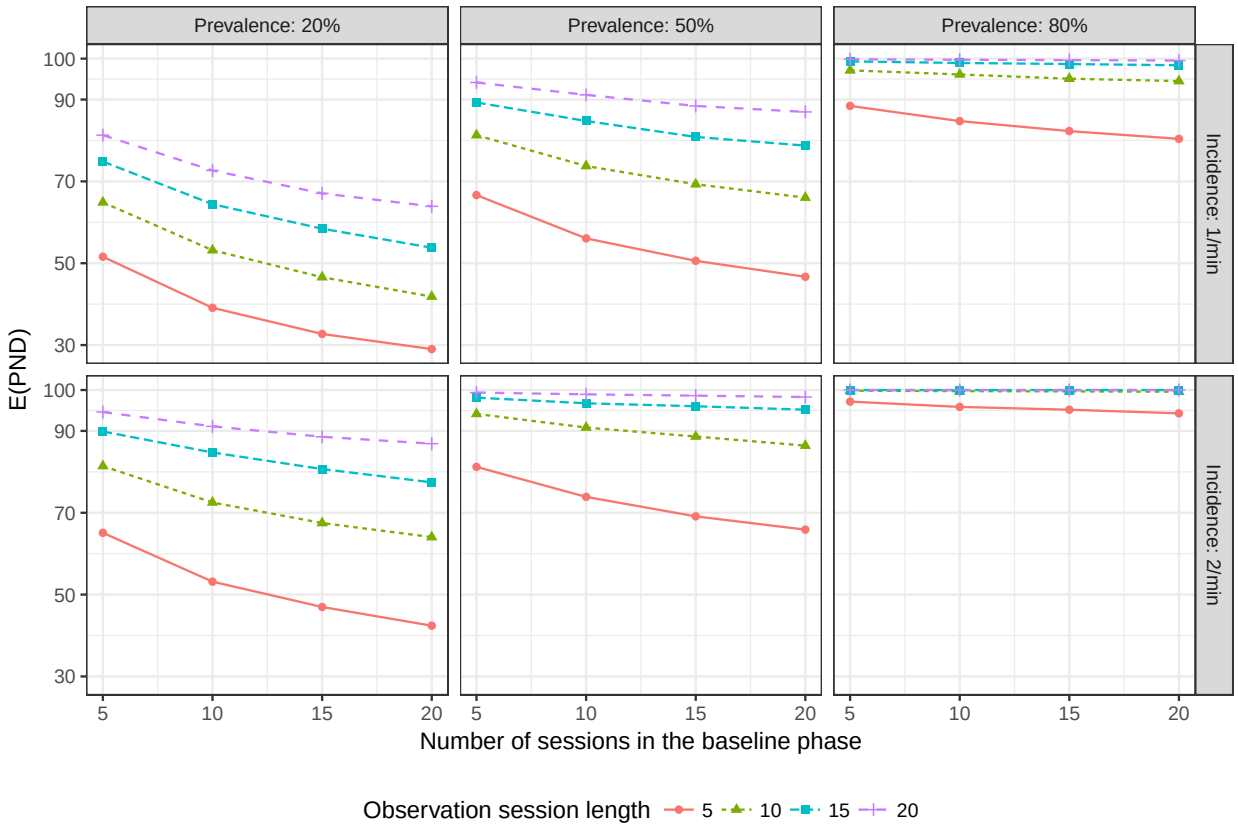

Figure 2. Expected value of PND for varying session lengths and vary numbers of observations in the baseline phase, when treatment leads to a 50% change and the outcome is measured using continuous recording.

(1994), who demonstrated that, when treatment has no effect, the magnitude of PND depends strongly on the number of sessions in the baseline phase.

When treatment produces beneficial effects, PND can become even more sensitive to the number of sessions in the baseline phase, while also becoming sensitive to both observation session length and recording system. Figure 2 provides a more detailed illustration of the degree to which PND is sensitive to the number of baseline observations and to session length, depicting the expected value of PND for a 50% change due to treatment, where the outcome is measured using continuous recording. When prevalence is high, PND is at or near ceiling across all of the variations in procedural factors. However,

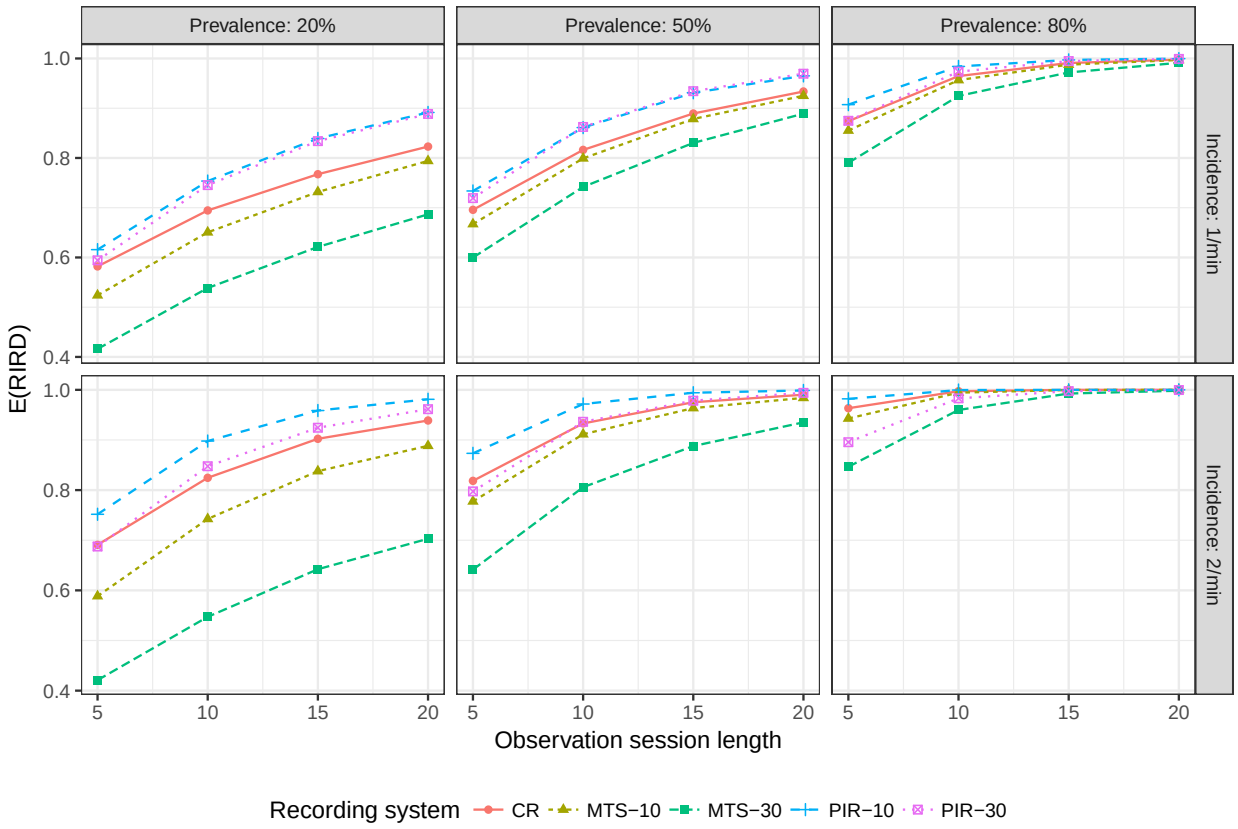

Figure 3. Expected value of RIRD for varying session lengths and recording systems, when treatment leads to a 50% change and both phases include 10 sessions. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

for lower levels of prevalence, PND is highly sensitive. For instance, when prevalence is 20% and incidence is twice per minute, the expected value ranges from 42% to 95%, depending on session length and the number of baseline observations. By the guidelines of Scruggs and Mastropieri (1998), the same intervention might appear to be "ineffective" or "very effective" depending solely on features of the study's design.

PAND and RIRD. Results for PAND and RIRD are presented in the second and third row of Figure 1. These two measures have quite similar properties. When treatment has no effect, the measures are sensitive to the number of sessions in both the baseline and treatment phase. However, the degree of sensitivity decreases when treatment produces

more beneficial effects; further details are available in the supplementary materials.

Both PAND and RIRD are sensitive to observation session length and recording system. Figure 3 illustrates the influence of these factors on the expected value of RIRD, focusing on the subset of results where both phases include 10 sessions and treatment leads to a 50% reduction in behavior. It can be seen that the magnitude of RIRD is strongly affected by observation session length and by recording system—particularly for lower values of prevalence. For prevalence of 80%, as well as for 80% reductions in behavior (not shown), the expected value of RIRD is somewhat constrained by ceiling levels, which restricts the degree of sensitivity to these factors. Results for PAND are similar; further details are available in the supplementary materials.

NAP. Conditional range distributions for NAP appear in the fourth row of Figure 1. The expected value of NAP is unaffected by the number of observations in the baseline phase or the treatment phase. Furthermore, the expected value of NAP is always exactly 50% when treatment has no effect on the behavior, regardless of the length of the observation sessions or the recording system used to collect outcome data. Consequently, these factors only matter when there is a non-null change in behavior due to treatment.

For non-null changes due to treatment, NAP is sensitive to observation session length and recording system. Figure 4 provides further detail regarding the degree of these sensitivities, plotting the expected value of NAP when treatment leads to a 50% decrease in behavior, for varying observation session lengths and recording systems; each panel displays results for a different combination of prevalence and incidence during baseline.[12] For some types of behavior, the magnitude of NAP is highly sensitive to the length of the observation session and to which recording procedure is used. For instance, when baseline prevalence is 20%, baseline incidence is twice per minute, and observation sessions are 10 min, using 30 s MTS leads to an expected value of 81% (a “medium” effect), whereas using

[12]When treatment leads to an 80% decrease in behavior, the expected value of NAP is at or near the ceiling level of 100% across most conditions in the simulation.

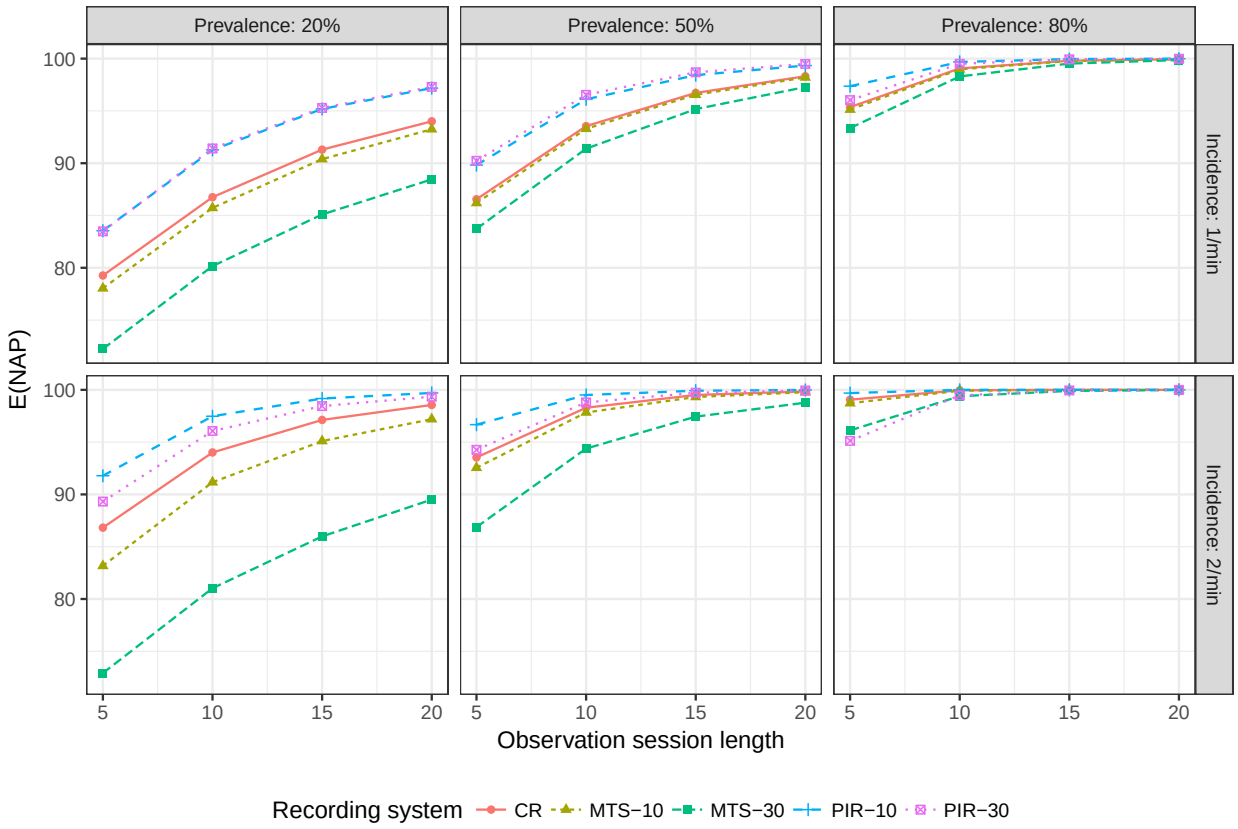

Figure 4. Expected value of NAP for varying recording systems and session lengths, when treatment leads to a 50% change in behavior. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

10 s PIR leads to an expected value of 97% (a "large" effect).

The extent to which NAP is sensitive to these procedural factors depends on the characteristics of the behavior. When the behavior has higher baseline prevalence or baseline incidence, NAP becomes less sensitive to observation session length and to recording procedure. However, this reduced sensitivity is largely due to the fact that NAP is at or near the ceiling level of 100% for all session lengths and recording systems. Thus, for changes in behavior in the range to which it is calibrated, the expected value of NAP is sensitive to the choice of observation session length and recording system.

PEM. The results for PEM (row 5 of Figure 1) are very similar to those for NAP. Like NAP, PEM is mostly unaffected by the number of observations in either phase and its expected value is 50% when treatment has no effect on the outcome. PEM is also sensitive to observation session length and recording procedure when the behavioral characteristics are in the range within which PEM can discriminate. The supplementary materials provide further details about the characteristics of PEM.

Parametric measures

Figure 5 depicts the conditional ranges for each of the parametric effect size measures, including the basic SMD (d), the bias-corrected SMD (g), the basic LRR (R_1), and the bias-corrected LRR (R_2), with respect to each of the four procedural factors. Its construction parallels that of Figure 1.

Standardized mean difference. Results in the first row of Figure 5 indicate that the basic SMD estimator (d) has a small-sample bias that induces sensitivity to the number of sessions in the baseline phase. However, the bias-corrected estimator (g) is only minorly affected by the number of sessions in each phase. The remaining discussion therefore focuses on g; the supplementary materials provide details about d.

The second row of Figure 5 indicates that the bias-corrected SMD is mostly unaffected by the number of baseline sessions, but does have large conditional range under a small set of conditions. These conditions all occur when prevalence is 80% and outcomes are recorded using 30-s PIR, which leads to measurements that are all very near ceiling levels. Under conditions where ceiling effects are not as severe, g is largely stable with respect to the number of observations in each phase. However, the index is strongly influenced by observation session length and recording system when treatment has non-null effects on behavior. To illustrate further, Figure 6 displays the expected value of g for varying session lengths and recording systems, based on the subset of results where treatment leads to a 50% reduction in behavior and both phases include 10 sessions. Within each panel, it can be seen that the magnitude of g is strongly influenced by observation

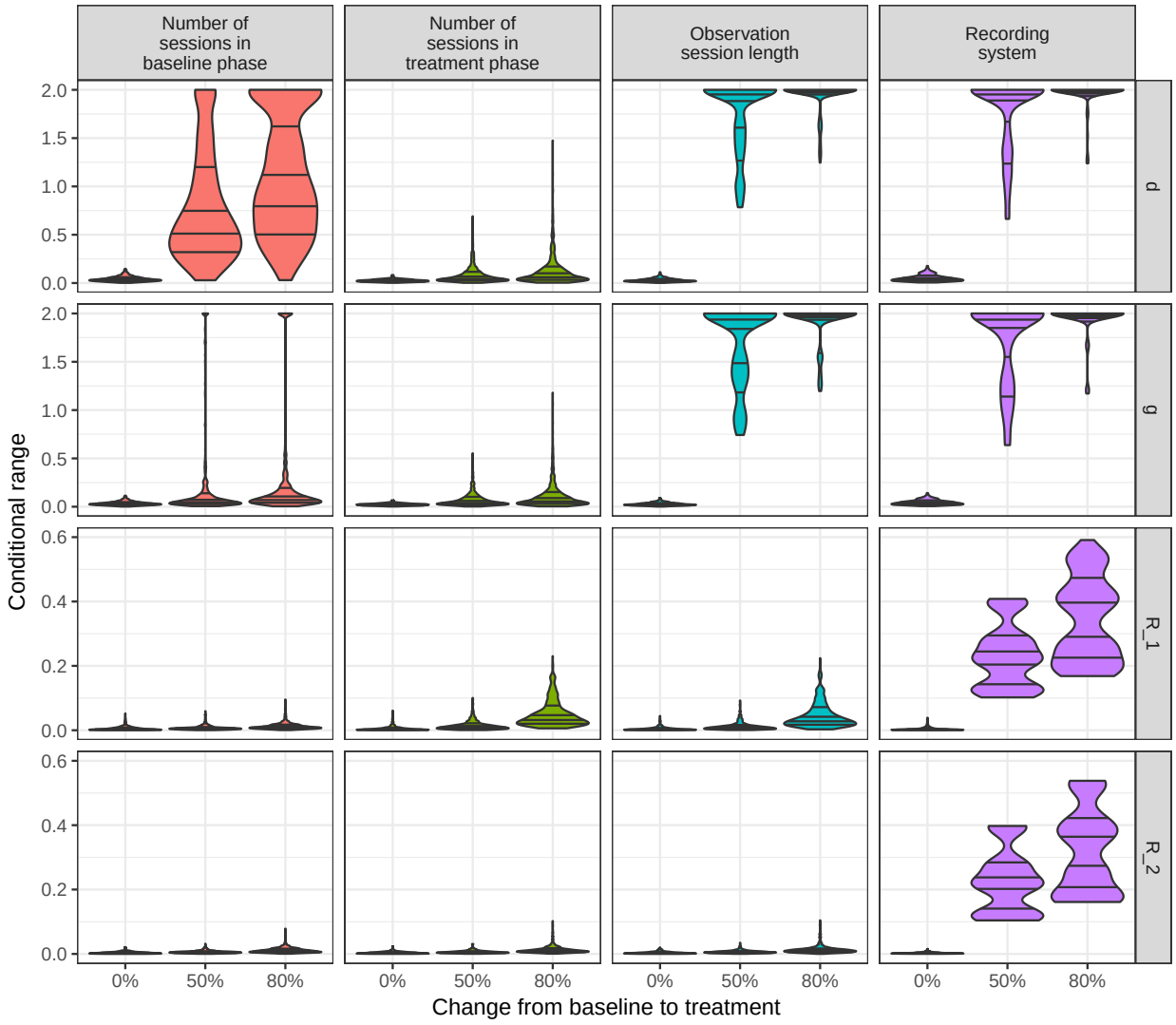

Figure 5. Conditional range distributions of the parametric effect size measures for each procedural factor, by percentage change from baseline to treatment. For clarity of illustration, the conditional ranges of d and g are truncated at 2.0.

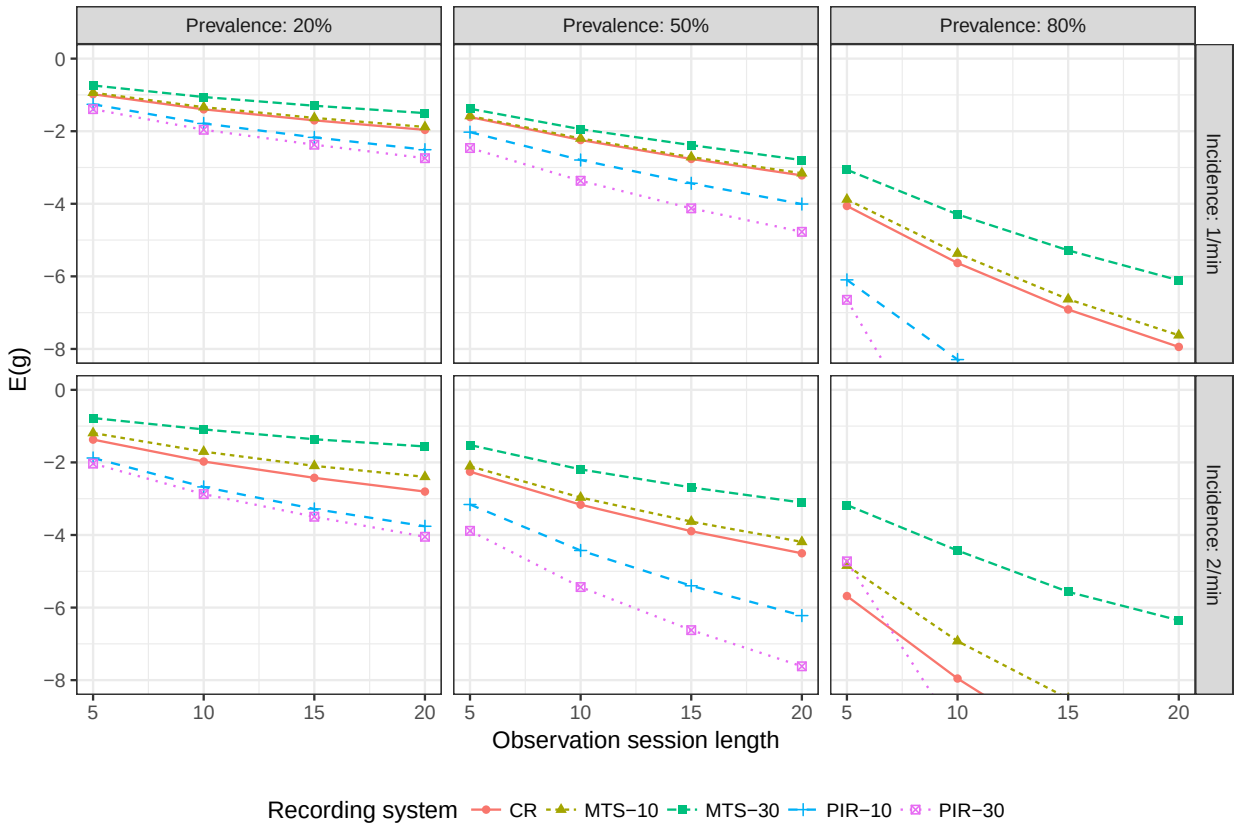

Figure 6. Expected value of bias-corrected SMD (g) for varying session lengths and recording systems, when treatment leads to a 50% change in behavior and both phases include 10 sessions. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

session length and recording system. For a given recording system, longer session lengths lead to effects that are larger in absolute magnitude, with differences between 10 min sessions and 20 min sessions exceed 0.5 SD under many conditions. Differences between recording systems are also large; for example, when prevalence is 50%, incidence is once per minute, and sessions are 10 min in length, the expected value of g varies by 0.59 depending on whether the outcome is measured using a 10 s MTS or 10 s PIR system.

Log response ratio. Figure 5 depicts results for the basic and bias-corrected LRR estimators in the third and fourth rows, respectively. Similar to the results for the SMD

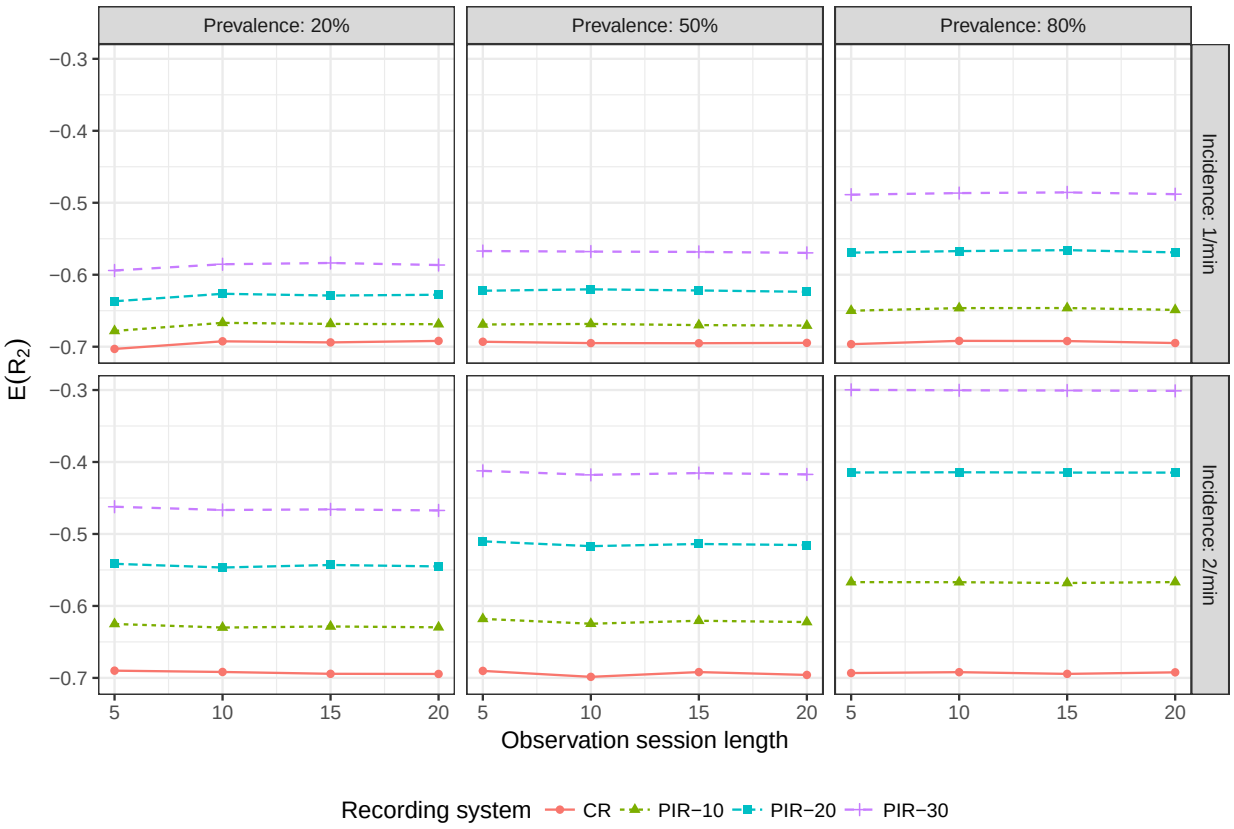

Figure 7. Expected value of R_2 for varying recording systems and session lengths, when treatment leads to a 50% change in behavior and each phase includes 10 sessions. CR = continuous recording; PIR = partial interval recording. Results for momentary time sampling systems are omitted because they are identical to results for continuous recording.

estimators, the basic moment estimator (R_1) has a small-sample bias, while the bias-corrected estimator (R_2) is only minorly influenced by the number of sessions in each phase. The expected value of R_1 is sensitive to the lengths of both the baseline phase and the treatment phase, particularly when treatment leads to larger reductions in behavior. Bias correction reduces this sensitivity, so that only very slight sensitivities remain under conditions where the outcome is measured imprecisely and when the phases include few sessions. The remaining discussion therefore focuses on R_2, while the supplementary materials provide further details about R_1.

Figure 5 indicates that—unlike any of the other effect size indices–the bias-corrected LRR is not affected by observation session length, although it is sensitive to the choice of recording system. To further illustrate the degree of sensitivity, Figure 7 depicts the expected value of R_2 for varying observation session lengths and recording systems, based on the subset of results where treatment leads to a 50% decrease in behavior and each phase includes 10 sessions. It can be seen that the expected value of R_2 is not affected by session length. Although its expected value remains stable across continuous recording and MTS systems, R_2 is biased towards zero when a PIR system is used, with longer intervals leading to a larger bias. In summary, the bias-corrected LRR is insensitive to the number of sessions in the baseline and treatment phases, insensitive to session length, and insensitive to the choice of interval length for MTS recording systems or to use of continuous recording; however, it is sensitive to the choice of interval length if the behavior is measured using a PIR system.

Discussion

Allison and Gorman (1994) first raised concerns about the procedural sensitivity of SCD effect size indices, demonstrating that the magnitude of PND is systematically affected by the number of baseline observations. Building on this earlier work, this study has examined the procedural sensitivity of an expanded selection of effect size indices, including some recently proposed indices and other well-known indices that are commonly used in syntheses of single-case research. A further contribution has been to examine a broader set of procedural factors, including observation session length and recording system, using a realistic model for systematic behavioral observation data collected in a SCD.

Results of the state behavior simulation demonstrated that the magnitudes of all but one of the effect size indices are influenced by arbitrary procedural details—likely selected by the researcher on the basis of feasibility and resource availability—rather than solely by the magnitude of change produced by an intervention. Several of the measures, including

PND, PAND, and RIRD, are sensitive to the number of sessions in the baseline or treatment phase. Other measures, including PEM, NAP, and the bias-corrected SMD, are not influenced by sample size but are sensitive to the length of the observation session and the recording system used to measure the behavior. Such procedural sensitivities represent an important limitation for the interpretation of these indices as effect sizes—and particularly for their use in meta-analysis—because they do not provide a fair basis for comparison across cases or studies that differ on procedural details.

Results of the event behavior simulation (reported in the supplementary materials) were broadly consistent with these findings. Findings from both simulations are summarized in Table 3. The main difference in findings concerns the degree to which the indices are sensitive to the choice of recording system. Across all effect sizes, the choice of recording system had only slight or moderate effects on magnitude when measuring event behaviors, whereas the choice of recording system generally had stronger effects for state behaviors. However, this finding may be limited by the range of conditions examined in the event behavior simulation. Other work, also based on the alternating renewal process model, has identified conditions under which the use of partial interval recording to measure event behavior can produce highly misleading inferences, such as concluding that treatment reduces the incidence of an undesirable behavior when in fact it increases it (Pustejovsky & Swan, 2015). Thus, the lack of sensitivity to recording system in the event behavior simulation might not hold more broadly.

For both state behaviors and event behaviors, the expected value of the bias-corrected LRR was stable across designs with varying numbers of observations in each phase, varying observation session lengths, and some (though not all) recording systems. This stands in contrast to the other indices and suggests that the LRR may be particularly appropriate and useful as an effect size for SCDs that use direct observation of behavior to measure the dependent variable. Because of its close connection with proportionate change, the LRR may also be intuitively appealing to applied behavioral researchers and clinicians,

Table 3

Procedural sensitivities of effect sizes for single-case designs

	Number of baseline sessions		Number of treatment sessions		Observation session length		Recording system	
Behavior	State	Event	State	Event	State	Event	State	Event
PND	X	X	-	-	X	X	X	-
PAND	X	X	X	X	X	X	X	-
RIRD	X	X	X	X	X	X	X	-
NAP	-	-	-	-	X	X	X	-
PEM	-	-	-	-	X	X	X	-
SMD	-	-	-	-	X	X	X	X
LRR	-	-	-	-	-	-	X	X

"X" indicates that an effect size is sensitive to a procedural factor. PND = percentage of non-overlapping data; PEM = percentage exceeding the median; PAND = percentage of all non-overlapping data; RIRD = robust improvement rate difference; NAP = non-overlap of all pairs; SMD = bias-corrected standardized mean difference; LRR = bias-corrected log response ratio.

who commonly conceptualize and discuss treatment impacts in terms of percentage change between phases (Campbell & Herzinger, 2010; Marquis et al., 2000).

A limitation of the LRR is that it remains sensitive to the use of partial interval recording systems of varying length. However, this may have less to do with the effect size measure than with the PIR system itself, which systematically over-estimates the prevalence of state behaviors, to an extent that depends both on interval length and on other features of the behavior (Kraemer, 1979; Wirth, Slaven, & Taylor, 2014). Another limitation of the LRR is that it does not account for time trends within the baseline or treatment phases. Developing methods for estimating LRRs while accounting for linear or non-linear time trends is an important goal for further research.

The present study examined the procedural sensitivity of effect size indices, while

holding constant the true effect magnitude. A related concern is the extent to which effect size indices can differentiate between effects vary in magnitude. Using a collection of 200 SCD series, Parker, Vannest, and Davis (2011) found that most NOMs could not discriminate between effects at larger magnitudes. In particular, all of the NOMs classified at least 10% of the series at ceiling levels, with PEM classifying over 50% of the series as having complete non-overlap. I observed similar behavior in the simulation studies based on the ARP model, in that several of the NOMs were at or near ceiling levels under some simulation conditions, which mitigated their procedural sensitivities (for example; NAP and PEM were always near ceiling for prevalence of 80%, incidence of of twice per minute, and a 50% reduction in behavior; see Figures 4 and S6). While this does support to the findings reported by Parker, Vannest, and Davis (2011), findings from the present simulation also point towards the need to take into account procedural differences between SCDs when investigating the correspondence between effect size indices and effects of varying magnitude.

The problems with the NOMs identified in this analysis add to a growing body of criticism of these measures. Researchers have criticized the NOMs because they do not align well with visual inspection of study results (Wolery et al., 2010), although other studies have reported moderate or strong agreement between some NOMs and visual analysis (Parker & Vannest, 2009; Petersen-Brown et al., 2012). Others have criticized the NOMs because they lack valid methods to quantify their sampling uncertainty (Shadish et al., 2008), which makes it difficult to apply conventional meta-analytic techniques for synthesis. At the same time, this analysis has demonstrated that procedural sensitivity is also a problem with the most commonly used parametric effect size, the within-case SMD. Of course, as a parametric effect size, the SMD and related approaches have an advantage that the distributional assumptions on which they are premised can be assessed in a given application. Thus, the findings of this analysis are consistent with the recognized need to further develop statistical methods for SCD data that are appropriate for outcomes

measured as counts or rates (e.g. Shadish, 2014b). This study also illustrates the importance of validating any such new developments using realistic data-generating models, such as the alternating renewal process. Furthermore, better statistical models and a stronger understanding of the psychometric properties of the types of outcome data used in single-case research would contribute to improved methods of synthesizing SCDs.

Limitations

The findings from this simulation study are limited in several respects. As in any simulation study, the findings are limited by the set of conditions examined, and any of the effect size indices may be less (or more) sensitive to operational features of the study design for patterns of behavior and measurement procedures outside of those considered here. More fundamentally, the findings hinge on the extent to which the alternating renewal process model is a plausible approximation to the real-life process of behavioral observation. Although special cases of the model have been used in a number of previous simulation studies of behavioral observation data (see references in Pustejovsky & Runyon, 2014), relatively little empirical data is available to investigate the model's distributional assumptions in detail. Until such evidence can be collected, the relevance of the model rests on its face validity, in that its formulation closely matches the physical process of collecting behavioral observation data.

The present study is also limited in that it focused on case-level effect sizes that are appropriate for use in the absence of time trends. Extant effect sizes that do account for time trends are all elaborations of the basic effect size indices, based on the same conceptualizations of treatment effect magnitude.[13] Thus, it is likely that they will exhibit procedural sensitivities that are similar to the basic indices from the same family of models. Further simulation work is warranted to verify this prediction and to better understand the properties and interpretation of these effect sizes. Similarly, it would be useful to investigate the procedural sensitivities of other effect size indices and

[13]Section S2 of the supplementary materials explains these relationships in detail.

meta-analysis techniques not reviewed here, including the between-case standardized mean difference (Shadish et al., 2014) and the multi-level meta-analysis models developed by Van den Noortgate and Onghena (2008).

Given that this study used simulation methods, a crucial avenue of further research is to examine the procedural sensitivities of SCD effect sizes using real data. An investigation of the associations between the magnitude of effect sizes and the operational characteristics of a set of real SCDs would provide an important source of empirical evidence regarding the issues identified in the present study. Based on analysis of a single SCD, Ledford (2015) reported preliminary evidence that use of interval-based recording systems can alter the magnitude of the NOMs, but further research is needed, ideally using a large corpus of data.

Finally, the scope of the simulation study was limited to four basic procedural features of SCDs, which are likely to vary across any collection of SCDs on a common topic. Collections of SCDs are also likely to exhibit variation in other procedural features, such as the operational definition of the focal behavior, the degree of inter-observer reliability achieved during direct observation, and the methods used to identify study participants (e.g., sampling for homogeneous versus heterogeneous cases). It is possible that the effect size indices examined in this study are sensitive to these other procedural features, and future research should investigate the extent of these procedural sensitivities using both simulation and empirical data. However, such investigations will likely be aided by using effect size indices, such as the LRR, that are relatively insensitive to other common dimensions of procedural variation (i.e., the basic features investigated this this study). Reducing extraneous variation in some dimensions will likely make it easier to identify other factors that explain variation in effect magnitude—whether such factors represent operational details or substantively important constructs.

Implications for applied research

In light of the procedural sensitivity of the NOMs and of the SMD, as well as other criticisms that have been raised about these measures, researchers interested in comparing or meta-analyzing evidence from SCDs with directly observed behavioral outcomes should exercise caution in interpreting them as effect sizes. Specifically, researchers should discontinue use of PND, PAND, and RIRD because they are all influenced by the number of observations in the baseline and/or treatment phase—procedural details that are quite likely to vary across cases, even within a single study (i.e., cases in a multiple baseline design necessarily have different numbers of observations in the baseline phase). Researchers who use the NAP, Tau, PEM, or bias-corrected SMD indices should be aware that, although not affected by the number of sessions per phase, the magnitude of these measures is affected by other details of the outcome measurement procedures. If a collection of SCDs to be meta-analyzed includes many studies that measured outcomes through direct observation, researchers should consider using the bias-corrected LRR because it is relatively unaffected by study procedures that are likely to vary across a collection of SCDs.

A further implication of this study is that systematic reviews of SCDs should devote more attention to the outcome measurement procedures and study designs on which their findings are based. In particular, systematic reviews should report details regarding the distribution of observation session lengths, recording systems, and the number of sessions per phase in the included studies. In addition to reporting descriptive information about the range of procedures used, meta-analyses of SCDs should investigate whether differences in outcome measurement procedures moderate the magnitude of effect sizes.

There is a long tradition of using non-overlap measures—and especially PND—to characterize the results of SCDs, which has continued despite stringent methodological critiques. Given this history, it seems likely that researchers conducting meta-analyses of SCDs might persist in reporting the most well-known indices as part of primary studies or

systematic reviews of single-case research. If these measures do continue to be widely reported, readers should be aware that these effect sizes can be sensitive—sometimes highly so—to procedural variation in study design and outcome measurement procedures. They should not be interpreted as pure measures of treatment effect magnitude.

References

Adamson, R. M., & Wachsmuth, S. T. (2014). A review of direct observation research within the past decade in the field of emotional and behavioral disorders. *Behavioral Disorders*, *39*(4), 181–189.

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

Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, *32*(8), 885–890. doi: 10.1016/0005-7967(94)90170-8

Ayres, K., & Gast, D. L. (2010). Dependent measures and measurement procedures. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 129–165). New York, NY: Routledge.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–236). New York, NY: Russell Sage Foundation.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: a quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *24*(2), 120–138. doi: 10.1016/S0891-4222(03)00014-3

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 417–450). New York, NY: Routledge.

Center, B. A., Skiba, R. J., & Casey, A. (1985-86). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4),

387–400. doi: 10.1177/002246698501900404

Conroy, M. A., Dunlap, G., Clarke, S., & Alter, P. J. (2005). A descriptive analysis of positive behavioral intervention research With young children with challenging behavior. *Topics in Early Childhood Special Education*, *25*(3), 157–166. doi: 10.1177/02711214050250030301

DiCarlo, C. F., Reid, D. H., & Stricklin, S. B. (2003). Increasing toy play among toddlers with multiple disabilities in an inclusive classroom: A more-to-less, child-directed intervention continuum. *Research in Developmental Disabilities*, *24*(3), 195–209. doi: 10.1016/S0891-4222(03)00025-8

Duan, N., Kravitz, R. L., & Schmid, C. H. (2013). Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8 SUPPL.8), S21–S28. doi: 10.1016/j.jclinepi.2013.04.006

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation*, *24*(3-4), 305–314. doi: 10.1080/09602011.2014.903198

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*(8), 761–8. doi: 10.1097/MLR.0b013e318215d90d

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders*, *37*(2), 55–77.

Gast, D. L. (2010). Applied research in education and behavioral sciences. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 1–19). New York, NY: Routledge.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science*, *20*(1), 71–79. doi: 10.1177/002188638402000113

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, *50*(2), 162–183. doi: 10.1080/00273171.2014.973989

Heath, A. K., Ganz, J. B., Parker, R. I., Burke, M., & Ninci, J. (2015). A meta-analytic review of functional communication training across mode of communication, age, and disability. *Review Journal of Autism and Developmental Disorders*, *2*(2), 155–166. doi: 10.1007/s40489-014-0044-3

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. doi: 10.3102/10769986006002107

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171. doi: 10.1111/j.1750-8606.2008.00060.x

Hedges, L. V., Gurevitch, J., & Curtis, P. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *35*(10), 2463–2476. doi: 10.1016/j.ridd.2014.06.017

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179. doi: 10.1177/001440290507100203

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage Publications, Inc.

Kraemer, H. C. (1979). One-zero sampling in the study of primate behavior. *Primates*, *20*(2), 237–244. doi: 10.1007/BF02373376

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: methodological and statistical advances* (pp. 91–125). Washington, DC: American Psychological Association.

Ledford, J. R. (2015). The effects of interval-based measurement on the estimation of effect sizes in single case research. In Ganz (Ed.), *Issues in and application of meta-analyses and syntheses of single-case experimental research in autism and developmental disabilities.* San Antonio, TX: Association for Behavior Analysis International Conference. Retrieved from `http://hdl.handle.net/1969.1/154270`

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications, Inc.

Ma, H.-H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification*, *30*(5), 598–617. doi: 10.1177/0145445504272974

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the special series: Issues and advances of synthesizing single-case research. *Remedial and Special Education*, *34*(1), 3–8. doi: 10.1177/0741932512466269

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology*, *49*(5), 529–54. doi: 10.1016/j.jsp.2011.05.001

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality*, *19*(2), 109–135. doi: 10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner,

R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*(3), 301–321. doi: 10.1016/j.jsp.2011.03.004

Maggin, D. M., Zurheide, J., Pickett, K. C., & Baillie, S. J. (2015). A systematic evidence review of the check-in/check-out program for reducing student challenging behaviors. *Journal of Positive Behavior Interventions*, *17*(4), 197–208. doi: 10.1177/1098300715573630

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*(4), 1262–1271. doi: 10.3758/BRM.41.4.1262

Marquis, J. G., Horner, R. H., Carr, E. G., Turnbull, A. P., Thompson, M., Behrens, G. A., ... Doolabh, A. (2000). A meta-analysis of positive behavior support. In R. Gersten, E. P. Schiller, & S. Vaughan (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues* (pp. 137–178). Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, S., Quinn, F., Vieira, R., O'Brien, N., White, M., Johnston, D. W., & Sniehotta, F. F. (2017). The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: a systematic literature overview. *Health Psychology Review*, *7199*, 1–17. doi: 10.1080/17437199.2017.1316672

Mudford, O. C., Locke, J. M., & Jeffrey, K. (2011). Rates of responding measured by continuous recording in applied behavioral research. *Behavioral Interventions*, *26*(1), 41–49. doi: 10.1002/bin.323

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the Journal of Applied Behavior Analysis (1995-2005). *Journal of Applied Behavior Analysis*, *42*(1), 165–169. doi: 10.1901/jaba.2009.42-165

Parker, R. I., Hagan-Burke, S., & Vannest, K. J. (2007). Percentage of all non-overlapping

data (PAND): An alternative to PND. *The Journal of Special Education*, *40*(4), 194–204. doi: 10.1177/00224669070400040101

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–67. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*(2), 135–150. doi: 10.1177/001440290907500201

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*(4), 303–22. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 127–151). Washington, DC: American Psychological Association. doi: 10.1037/14376-005

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*(2), 284–299. doi: 10.1016/j.beth.2010.08.006

Petersen-Brown, S., Karich, A. C., & Symons, F. J. (2012). Examining estimates of effect using non-overlap of all pairs in multiple baseline studies of academic intervention. *Journal of Behavioral Education*, *21*(3), 203–216. doi: 10.1007/s10864-012-9154-0

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*(3), 342–359. doi: 10.1037/met0000019

Pustejovsky, J. E. (2016). *ARPobservation: Simulating recording procedures for direct observation of behavior.* R package Version 1.1. Retrieved from `http://cran.r-project.org/web/packages/ARPobservation`

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. doi: 10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., & Ferron, J. M. (2016). Research synthesis and meta-analysis of single-case designs. In J. M. Kaufmann, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education, 2nd edition.* New York, NY: Routledge.

Pustejovsky, J. E., & Runyon, C. (2014). Alternating renewal process models for behavioral observation: Simulation methods, software, and validity illustrations. *Behavioral Disorders*, *39*(4), 211–227.

Pustejovsky, J. E., & Swan, D. M. (2015). Four methods for analyzing partial interval recording data, with application to single-case research. *Multivariate Behavioral Research*, *50*(3), 365–380. doi: 10.1080/00273171.2015.1014879

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, *16*(3), 157–252. doi: 10.2307/1165191

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification*, *22*(3), 221–242. doi: 10.1177/01454455980223001

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*(1), 9–19. doi: 10.1177/0741932512440730

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–43. doi: 10.1177/074193258700800206

Shadish, W. R. (2014a). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*, *52*(2), 109–122. doi: 10.1016/j.jsp.2013.11.009

Shadish, W. R. (2014b). Statistical analyses of single-case designs: The shape of things to

come. *Current Directions in Psychological Science*, *23*(2), 139–146. doi: 10.1177/0963721414524773

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*(2), 123–147. doi: 10.1016/j.jsp.2013.11.005

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*, *2*(3), 188–196. doi: 10.1080/17489530802581603

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. doi: 10.3758/s13428-011-0111-y

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. doi: 10.1037/a0029312

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education*, *24*(4), 438–458. doi: 10.1007/s10864-015-9221-4

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, *41*(4), 427–467. doi: 10.1177/0145445516676750

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The design, conduct and report of single-case research: Resources to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation*, *24*(3-4), 315–331. doi: 10.1080/09602011.2013.875043

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication*

Assessment and Intervention, *2*(3), 142–151. doi: 10.1080/17489530802505362

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. doi: 10.2307/1165329

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment*, *11*(3), 281–296.

White, O. R. (1987). Some comments concerning "The quantitative synthesis of single-subject research". *Remedial and Special Education*, *8*(2), 34–39. doi: 10.1177/074193258700800207

Wirth, O., Slaven, J., & Taylor, M. A. (2014). Interval sampling methods and measurement error: A computer simulation. *Journal of Applied Behavior Analysis*, *47*(1), 83–100. doi: 10.1002/jaba.93

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi: 10.1177/0022466908328009

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., ... Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders*, *45*(7), 1951–1966. doi: 10.1007/s10803-014-2351-z