Operational sensitivities of non-overlap effect sizes for single-case designs

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.

Operational sensitivities of non-overlap effect sizes for single-case designs

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; SCDs are therefore true experiments in the sense of Shadish, Cook, and Campbell (2002). Changes in the pattern of outcomes during the phases where the intervention is present versus where it is absent are taken as evidence that the intervention has a causal effect (i.e., a functional relationship with the outcome) for that individual. In the logic of single-case research, systematic evidence for an effect accumulates through replication of the pattern at several times, across individuals, settings, or target outcomes (Horner & Odom, 2014).

There is a long-recognized need for principled methods of synthesizing data from SCDs. At a time when many other fields were just beginning to attend to meta-analysis, Gingerich (1984) argued that meta-analytic methods could be used to draw more generalized inferences than would be warranted from single-case studies considered separately, and could provide a means of identifying important sources of variation in intervention effects.Allison and Gorman (1993) and others argued for the need to include evidence from SCDs in comprehensive syntheses of intervention effects. More recently, the turn towards evidence-based practice and policy-making has led to renewed attention to methods for synthesizing SCDs. Several prominent research organizations, including Divisions 12 and 16 of the American Psychological Association (Chambless & Hollon, 1998; Chambless & Ollendick, 2001; Kratochwill & Stoiber, 2002), the Council for Exceptional Children's Division for Research (Horner et al., 2005; Odom et al., 2005), and the Institute for Education Sciences' What Works Clearinghouse initiative (Kratochwill et al., 2012), have proposed guidelines for establishing evidence-based practices on the basis of evidence from SCDs. While these guidelines make clear that

data from SCDs should be weighed when evaluating the evidence base, many questions remain about how best to do so.

One of the basic decisions that must be made in any quantitative synthesis of intervention research is what effect size to use to quantify the magnitude of treatment effects. Ideally, an effect size measure should be on a scale that is easy to interpret and that also allows for comparison with other studies in the same area (Hedges, 2008; Lipsey & Wilson, 2001). Consequently, a good effect size measure should be relatively insensitive to incidental features of a study's design, such as the sample size or the choice of measurement procedures. Effect sizes metrics that are sensitive to such operational details will tend to obscure substantive variation in study results and reduce the interpretability of the results of a synthesis.

A wide array of effect sizes have been proposed for use with SCDs (for an extensive review, see Beretvas & Chung, 2008), but there remains considerable disagreement regarding their merits (Shadish, Rindskopf, & Hedges, 2008). The most widely used effect size measures in single-case research are in the family of non-overlap statistics, which includes the percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), the non-overlap of all pairs (NAP; Parker & Vannest, 2009), and a number of others. These measures are sometimes described as non-parametric effect sizes, in that they are not premised on assumptions regarding the normality of outcome measures and are insensitive to outliers (Parker, Vannest, & Davis, 2011). These properties are seen as advantageous because many outcome measures used in single-case research are not well-modeled by normal distributions. However, little previous research has examined the characteristics of non-overlap effect sizes under data-generating models that are more plausible for the types of outcome measures used in single-case research. The present study aims to fill that gap, by studying the behavior of non-overlap effect sizes using data simulated from a realistic model for direct obervation of behavior.

Behavioral measures derived from systematic direct observation are by far the most common type of outcomes in single-case research (Gast, 2010). A variety of different procedures are used to record direct observation of behavior, including continuous recording, momentary time sampling, frequency counting, and partial interval recording. Each of these procedures can be used for shorter or longer periods of observation, and the number of observation sessions per phase may also vary from study to study. Thus, in order to synthesize SCDs that use behavioral outcome measures, an effect size metric is needed that is insensitive to variation in the outcome measurement procedures.

In order to study the properties of non-overlap metrics 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 model (Pustejovsky, 2014b; Rogosa & Ghandour, 1991), special cases of which have been used in many previous simulation studies of behavioral observation procedures (Pustejovsky & Runyon, 2014). The alternating renewal process is a model for the stream of behavior as it is actually perceived during an observation session. The model describes this behavior streams in terms of the duration of individual episodes of behavior and the lengths of time in between episodes of behavior. The alternating renewal process works by treating the episode durations and the interim times as random quantities, drawn from specified probability distributions. A key benefit of using the alternating renewal process model is that it mimics the actual process of observing a behavior in real time and recording data as one does so; as a result, the model captures the distinctive features of real, empirical behavioral observation data.

Using the alternating renewal process model, this study investigates the extent to which extant non-overlap measures of effect size are sensitive to procedural features of a single-case study, including the number of sessions during baseline and during treatment, the length of the observation sessions during which outcome measures are collected, and the procedure used to record behavioral observations. The analysis is organized as

follows: the next section reviews non-overlap statistics that have been proposed as effect size measures for SCDs; the following section explains the design of the simulation study; and the penultimate section describes the results. A brief conclusion discusses implications for synthesis of single-case research.

Non-overlap measures

This section briefly reviews several of the non-overlap statistics that have been proposed as effect size measures for SCDs.[1] Extant definitions of some of the non-overlap measures are ambiguous, which creates the possibility that different analysts might arrive at different values for the measures even when working with a common set of data. In order to be entirely precise and to ensure the reproducibility of our own analysis, I provide exact mathematical expressions for each of the measures. For sake of simplicity, I limit consideration to statistics that are appropriate for data that do not display time trends. Extensions of certain the non-overlap statistics have been proposed that accomodate time trends in the outcome measures (e.g. Parker, Vannest, Davis, & Sauber, 2011); however, the properties of these statistics largely resemble the properties of the analogous statistics that do not account for trends.

Each of the non-overlap statistics is defined in terms of a single contrast between a baseline phase and an intervention phase. Thus, it will suffice to consider a simple AB-type single-case design, with 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. Throughout, I assume that the dependent variable is defined in such a way that larger values correspond to a more beneficial outcome, so that increase in the outcome measure are desirable.[2] Denote the outcome measurements during the baseline phase as $Y_1^A, ..., Y_m^A$ and the outcome measurements

[1] Parker, Vannest, and Davis (2011) gives a more expansive reviews of non-overlap measures that includes examples demonstrating how to calculate them based on graphed data.

[2] For outcomes where an increase is beneficial, one would first multiply the outcome by -1 and then evaluate the specified formula for the non-overlap measure.

during the treatment phase as $Y_1^B, ..., Y_n^B$. Let $I(A)$ denote the indicator function, which is equal to one when A is true and equal to zero when A is false.

Percentage of non-overlapping data

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

$$\text{PND} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} I\left(Y_i^B > \max\left\{Y_1^A, ..., Y_m^A\right\}\right). \quad (1)$$

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 interpretted 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).

Since it was first proposed, PND has been widely criticized (e.g. Shadish et al., 2008; White, 1987; Wolery, Busick, Reichow, & Barton, 2010). In an analysis similar to the simulations presented in a later section, Allison and Gorman (1994) pointed out that the expected value of the PND statistic—that is, its average value across repeated samples—is strongly influenced by the number of observations in the baseline phase. In particular, longer baseline phases will tend to result in smaller values of PND, even when the intervention has no effect at all. They argued that this dependence on sample size makes the statistic unsuitable for use as an effect size metric. Others have pointed out that PND is highly sensitive to outliers because it uses the maximum value in the baseline phase as the basis for comparison with the treatment phase.Despite such objections, PND remains by far the most commonly applied effect size for synthesis of SCDs (Maggin, O'Keeffe, & Johnson, 2011; Scruggs & Mastropieri, 2012).

Percentage exceeding the median

In order to remedy some of the problems with the widely used PND statistic, Ma (2006) proposed an alternative that uses the median of the baseline phase (instead of the maximum) as the basis for comparison with the treatment phase. 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. 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 observation. Mathematically,

$$\text{PEM} = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\left[I\left(Y_i^B > M^A\right) + 0.5I\left(Y_i^B = M^A\right)\right], \tag{2}$$

where $M^A = \text{median}\left\{Y_1^A, ..., Y_m^A\right\}$. Like PND, PEM ranges in principle from 0 to 100%. Unlike PND, the expected magnitude of PEM is stable when the intervention 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 should be 50%. To our knowledge, no explicit guidance has been offered regarding what constitudes a small, medium, or large value of PEM.

Percentage of all non-overlapping data

Parker et al. (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 highest measurement in the baseline phase is less than the lowest measurement in the treatment phase. However, more recent descriptions of PAND use a subtly distinct definition, defining the statistic as the percentage of the total number of observations remaining after removing the minimum number of observations from either phase such that the highest remaining measurement from the baseline phase is less than

the lowest remaining measurement from the treatment phase (Parker, Vannest, & Davis, 2011, 2014). I employ the latter definition on the assumption that it supercedes the former.

PAND is rather difficult to express mathematically. Let $Y^A_{(1)}, Y^A_{(2)}, ..., Y^A_{(m)}$ denote the values of the baseline phase data, sorted in increasing order, and let $Y^B_{(1)}, Y^B_{(2)}, ..., Y^B_{(n)}$ denote the values of the sorted treatment phase data. In general, PAND can be calculated as

$$\text{PAND} = 100\% \times \frac{1}{m+n} \max\left\{(i+j)\, I\left(Y^A_{(i)} < Y^B_{(n+1-j)}\right)\right\}, \tag{3}$$

where the maximum is taken over the values $1 \leq i \leq m$ and $1 \leq j \leq n$.

Unlike PND and PEM, the logical range of the PAND statistic is not obvious. It is clear that 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 $\max\{m, n\}/(m+n)$, the number of observations in the longer of the two phases, divided by the total number of observations.[3] Parker et al. (2007); Parker, Vannest, and Davis (2011) indicated that 50% is the expected magnitude of PAND when the intervention has no effect on the outcome. This is clearly incorrect because, when m is not equal to n, the minimum possible value is larger than 50%. The simulation study described below demonstrates that, when the intervention has no effect, the expected magnitude of PAND depends on the lengths of each phase.

Phi/IRD

In addition to PAND, Parker et al. (2007) also proposed to use the Pearson's phi coefficient corresponding to a 2×2 table arrangement of the numbers obtained in

[3]Consider data in which the minimum of the baseline phase observations is larger than the maximum of the treatment phase observations. In order to obtain no overlap, one must either remove all baseline phase observations or all treatment phase 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.

calculating PAND, arguing that phi offers the advantage of having a known sampling distribution. In later work, the same authors instead suggest using a "robust" version of the phi coefficient where overlapping observations are evenly divided between the lower left and upper right cells of the 2×2 table, so that the row- and column-margins of the table are equal. In other work, Parker, Vannest, and Brown (2009) describe the "robust improvement rate difference" (IRD), which is exactly equivalent to the robust phi coefficient (Parker, Vannest, & Davis, 2011). I therefore focus on the "robust" IRD measure.

Following the examples provided in Parker, Vannest, and Davis (2011); Parker et al. (2014), robust IRD is calculated as follows. Let x be the minimum number of observations that must be removed from either phase so that the maximum of the remaining baseline phase observations is less than the minimum of the remaining treatment phase observations. Mathematically, $x = (m + n)(1 - \text{PAND}/100\%)$ where PAND is calculated as in Equation (3). IRD is then calculated as

$$\text{IRD} = \frac{n - x/2}{n} - \frac{x/2}{m}. \tag{4}$$

Staight-forward algebraic manipulations lead to the fact that IRD (robust phi) is a linear re-scaling of PAND, where

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

Note that the logical range of IRD therefore depends on the ratio of m to n. As with PAND, the expected magnitude of IRD when the intervention has no effect on the outcome is unclear, and in fact may depend on features of the study design. Despite this ambiguity, Parker et al. (2009) provided tentative benchmarks for the interpretation of its magnitude based on a comparison between the IRD statistics and expert visual assessments, 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 (p. 147).

Non-overlap of all pairs

Parker and Vannest (2009) proposed the non-overlap of all pairs (NAP) statistic and argued that it has several advantages over other non-overlap measures of effect size, including PND, PEM, and PAND.NAP involves comparisons between each point in the treatment phase and each point in the baseline phase, or a total of $m \times n$ pairs of comparisons. The statistic is defined as the percentage of all pairwise comparisons where the measurement from the treatment phase exceeds the measurement from the baseline phase (Parker & Vannest, 2009). Pairs of data points that are exactly tied are counted with a weight of 0.5. Mathematically,

$$\text{NAP} = 100\% \times \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[I\left(Y_j^B > Y_i^A\right) + 0.5 I\left(Y_j^B = Y_i^A\right) \right]. \tag{5}$$

Parker and Vannest (2009) noted that the numerator of NAP corresponds to the U statistic from the Wilcoxon rank-sum test. The logical range of NAP is from 0 to 100%, with a stable expected magnitude of 50% when the intervention has no effect on the outcome. Based on visual assessment of a corpus of SCD studies, Parker and Vannest (2009) characterized NAP values between 0 and 65% as "weak," values between 66% and 92% as "medium," and values between 93% and 100% as "large" (p. 364).

Tau

Parker, Vannest, Davis, and Sauber (2011) described the Tau effect size, which is closely related to NAP, but is designed to handle time trends in the baseline and treatment phases. In the absence of time trends, Tau is equal to the Spearman rank-correlation between the outcome measures and a binary variable indicating the treatment phase (taking the baseline phase as the reference category). It follows that, in the absence of time trends, Tau is simply a linear re-scaling of NAP:

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

Tau is therefore on a scale of -1 to 1, with an expected magnitude of 0 if the intervention has no effect on the outcome. Based on this simple relationship between Tau and NAP,

it seems reasonable to apply the benchmark values for NAP proposed by Parker and Vannest (2009), interpreting a value less than .30 as a "weak" effect, values between .30 and .84 as a "medium" effect, and a value greater than .84 as a "large" effect. Because NAP and Tau are so closely related, our simulation focuses on the former measure only.

Simulation design

I constructed a simulation study to examine the extent to which the non-overlap measures are sensitive to procedural features that are likely to vary across a collection of SCDs to be synthesized. This section describes the data-generating model, design, and procedures employed in the simulation study.

Data-generating model

The simulation uses the alternating renewal process model to generate realistic behavioral observation data. This model works by directly simulating behavior streams, each of which consists of episodes of behavior seperated by spans of time without behavior. The length of each behavioral episode is drawn at random from a distribution with a specified mean, and the length of each interim time is drawn at random from a different distribution with a specified mean. The process is repeated–alternately drawing behavioral episodes and interim times–until the sum of the episode lengths and interim times meets or exceeds the length of the observation session, denoted L. After a behavior stream has been simulated, a specified recording procedure can be used to derive summary measurements.[4] Because the behavior stream itself is randomly generated, the summary measurements exhibit random variation as well.

Several further details regarding the data-generating process must be specified in order to make the alternating renewal process fully operational. In an alternating renewal process, the main behavioral characteristic of interest is the prevalence, or the overall proportion of time that behavioral episodes occur. Behavior streams also vary in

[4]For further details about the recording procedures and summary measurements, see Pustejovsky and Runyon (2014).

their incidence, or the rate at which new behavioral episodes occur. Prevalence, incidence, or both prevalence and incidence may change between phases. The prevalence and incidence of the behavior within a given phase determine the mean event duration and mean interim time used to generate behavior streams. Additionally, the properties of the behavior stream will also depend to some extent on the parametric form of the distributions used to simulate event durations and interim times. Thus, to fully operationalize the model, values for prevalence and incidence in each phase of the design must be selected, as well as distributional forms for event durations and interim times.

Little systematic guidance exists regarding realistic values for the parameters of the behavior stream. Lacking strong empirical evidence, I used values for prevalence and incidence chosen to represent a range of conditions, and then verified the plausibility of the conditions by visual inspection of the simulated data. The prevalence of the behavior during the baseline phase was set to 20%, 50%, or 80% in order to capture a range of different types of behavior, such as mild, moderate, or severe problem behavior. Regarding the incidence of the behavior, **?** reported that SCDs published in the *Journal of Applied Behavior Analysis* between 1998 and 2007 displayed a median average 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 these findings, I set the incidence of the behavior during the baseline phase to once per minute or twice per minute. Based on our own experience, the majority of SCDs focus on behaviors in which a decrease is desiable; I therefore simulated data in which the treatment reduces the prevalence and incidence of the behavior by a certain percentage. I varied the change in behavior between zero, representing no effect of the intervention, and 80%, representing a fairly large decrease in the behavior.

Figure 1 displays simulated SCDs based on each combination of prevalence, and incidence, and change in behavior. The observations are generated using continuous recording with $L = 5$ min observation sessions, and $m = n = 10$ sessions are used in each

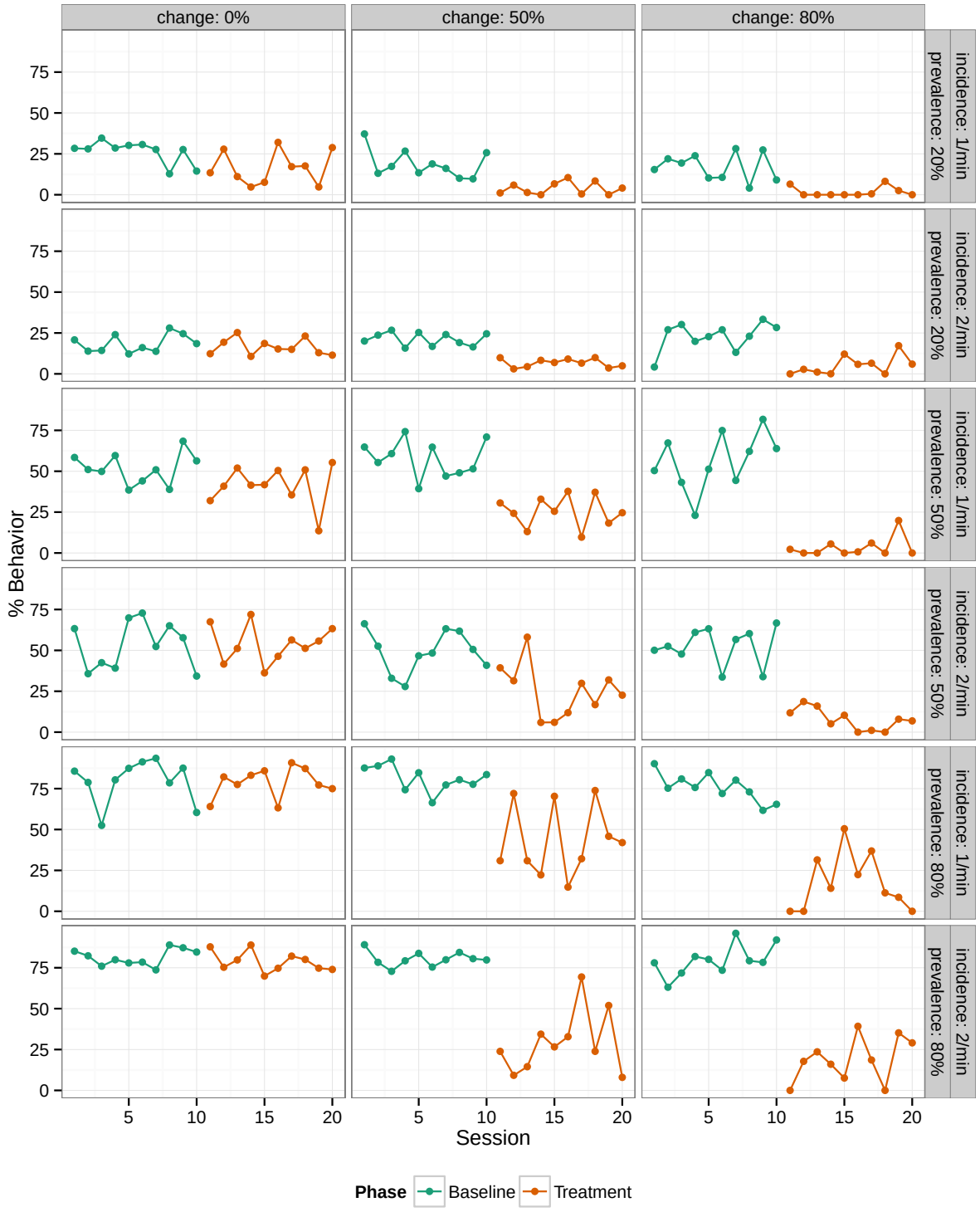

Figure 1. Simulated SCDs based on the alternating renewal process, for different levels of prevalence, incidence, and change in behavior.

phase. The reader may judge for themselves whether the simulated data appear to resemble the data from real SCDs.

Procedural factors

The simulation examined the following three procedural factors: recording procedure (including various interval lengths for intermittent recording procedures), length of observation session, and the number of observations in the baseline and treatment phases. In order to test the non-overlap measures under realistic conditions, I selected levels for these factors that resemble as closely as possible the procedures used in empirical SCDs.

Three procedures for recording direct observation of behavior are continuous recording (CR), momentary time sampling (MTS), and partial interval recording (PIR) (Gast, 2010). Reviews of the single-case literature indicate that all three of these procedures have been and continue to be used in practice (Adamson & Wachsmuth, 2014; Kelly, 1977; Mudford, Taylor, & Martin, 2009; Rapp et al., 2007). For example, a recent synthesis of SCDs examined the effect of functional behavior assessment interventions on student problem behavior (Gage, Lewis, & Stichter, 2012); of approximately 200 cases included in the synthesis, 65% were measured using PIR, 6% using CR, and 7% using MTS. MTS and PIR are both intermittent recording procedures, which involve making observations in each of many short intervals of time. 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 procedures may be used for longer or shorter observation sessions. For example, cases in the synthesis of functional behavior assessment interventions (Gage et al., 2012) were observed for between 5 and 60 min per session; 75% of cases were observed for 20 min or less. To emulate conditions typically used in practice, the simulation examined session lengths of 5, 10, 15, and 20 min.

Finally, SCDs use a wide range of phase lengths, with some phases consisting of

Table 1

Simulation design

Parameter	Levels
Prevalence	20%, 50%, 80%
Incidence (per min)	1, 2
Change (decrease)	0%, 50%, 80%
Recording procedure	CR, MTS (10, 20, 30 s), PIR (10, 20, 30 s)
Session length (L min)	5, 10, 15, 20
Baseline length (m)	5, 10, 15, 20
Treatment length (n)	5, 10, 15, 20

fewer than 5 observation sessions while others including far more. In a review of over 400 SCDs published between 2000 and 2010, Smith (2012) found that the average length of baseline phases was 10.2, 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 phases in included studies used initial baselines of 5 or more observations. The What Works Clearinghouse (WWC) standards for SCDs require that a design include at least 5 observations per phase in order to meet standards without reservations Kratochwill et al. (2012). The simulations were limited to designs that would meet the WWC standards in this respect, ensuring that the results would apply to SCDs that are considered to be of high quality. The simulations examined designs with 5, 10, 15, or 20 observations in the baseline phase, and 5, 10, 15, or 20 observations in the treatment phase. Both balanced and unbalanced designs were examined by including all 16 possible combinations of baseline lengths and treatment lengths.

Procedures

Table 1 summarizes the factors varied in the simulation study, which used a $3 \times 2 \times 3 \times 7 \times 4 \times 4 \times 4$ factorial design. The simulation was carried out using the ARPobservation package (Pustejovsky, 2014a) for the R statistical computing environment (R Core Team, 2014). For each combination of factor levels, I generated 10000 simulated AB designs and calculated the PND, PAND, IRD, PEM, and NAP statistics. I then averaged the simulated values across replications in order to estimate the expected magnitude of each statistic.

Results

This section presents the results of the simulation study for each of the non-overlap measures. I present the results in the form of figures that illustrate the degree to which the expected magnitude of each measure varies as a function of the procedural aspects of the design (i.e., recording procedure, session length, baseline length, and treatment length). For measures whose expected magnitude is known to be independent of the phase lengths, results are averaged across the levels of m and n. For clarity of presentation, some of the figures present results for selected subsets of the simulation conditions; in these cases, the results that are presented are generally consistent with the other simulation conditions.

PND

Figure 2 plots the expected magnitude of PND when the intervention has no effect, for the subset of results where $L = 5$ and for varying baseline lengths. Consistent with the findings of Allison and Gorman (1994), it can be seen that the null magnitude depends strongly on the number of observations in the baseline phase. For continuous recording data, where there is a negligible probability of two measurements being exactly equal, the expectation is exactly equal to

$$\mathrm{E}\left(\mathrm{PND}\right) = \frac{100\%}{m+1}$$

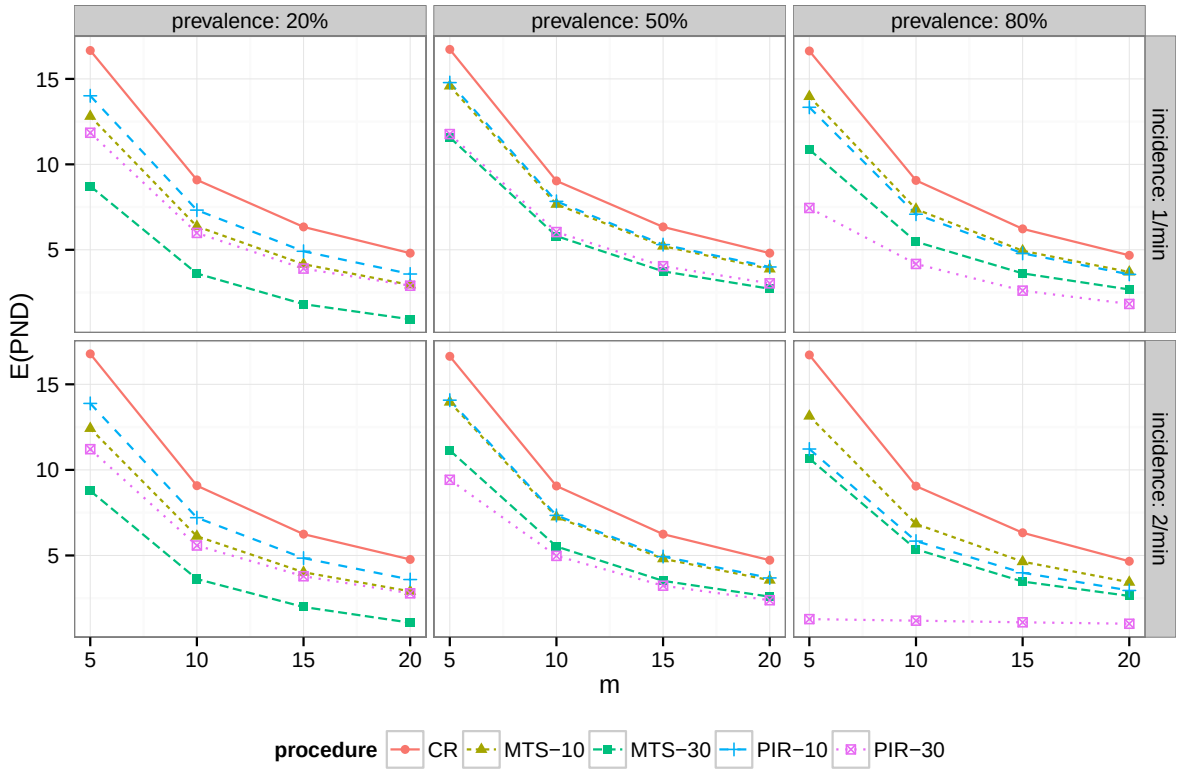

Figure 2. Expected magnitude of PND when the intervention has no effect, for $L = 5$ and varying session lengths.

across all values of L, prevalence, and incidence.[5] The expectation based on the other recording procedures differs from that for continuous recording only due to the possibility of exact ties.

In the conditions where the intervention produces beneficial effects, PND remains sensitive to baseline length and recording procedure and also becomes sensitive to length of the observation session. To illustrate the degree of sensitivity, Figure 3 plots the expected magnitude of PND for a 50% change due to intervention, where the outcome is measured using continuous recording (results for other recording procedures display a similar degree of sensitivity). When prevalence is high, PND is at or near ceiling across

[5]Allison and Gorman (1994) gave the approximate formula $\mathrm{E}(\mathrm{PND}) \approx 100\% \times (1 - 2^{-m})$, which is slightly less than the exact value.

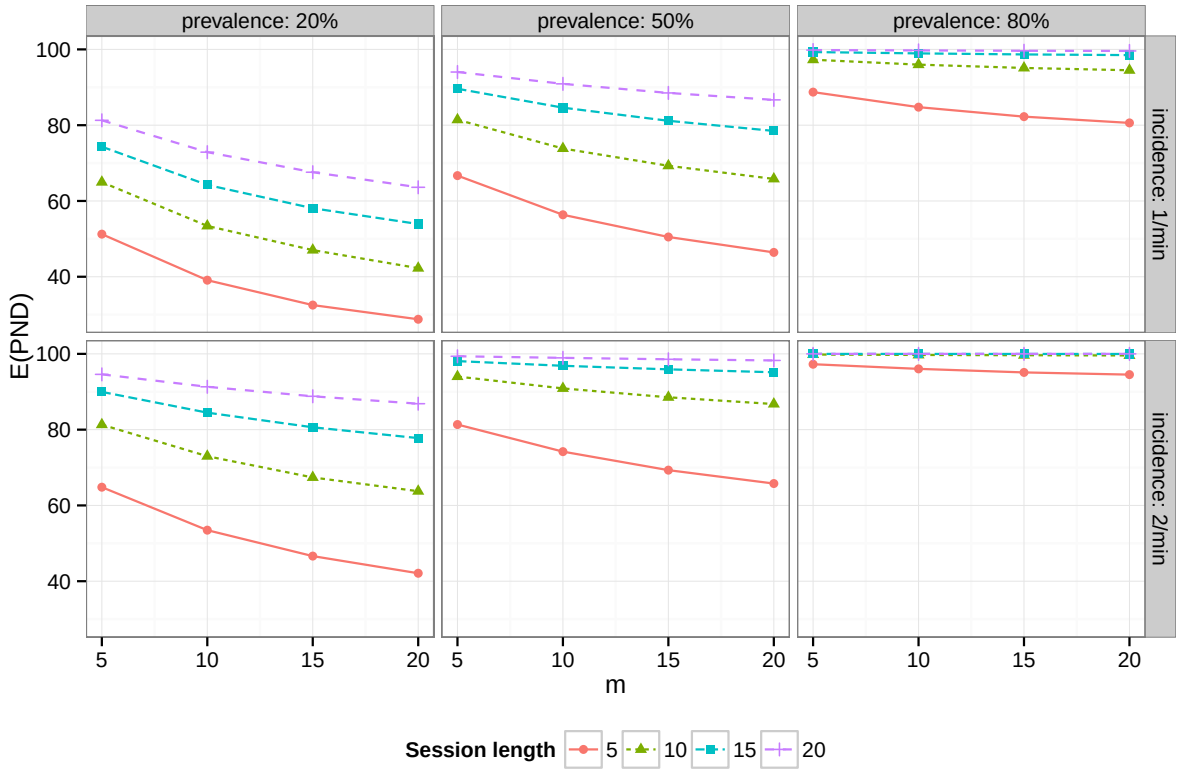

Figure 3. Expected magnitude of PND based on continuous recording data when the intervention leads to a 50% change, for varying session lengths.

all of the variations in procedural factors. However, for lower levels of prevalence, PND is highly sensitive to the procedural factors; for instance, when prevalence is 20% and incidence is twice per minute, the expectation ranges from 42 to 95, depending on the values of m and L.

PAND

Figure 4 depicts the expected magnitude of PAND as a function of the number of observations in the baseline phase (n) and in the intervention phase (m), for the subset of results where continuous recording is used for $L = 5$ minute sessions and where incidence is once per minute. Although Parker et al. (2007); Parker, Vannest, and Davis (2011) suggested that 50% is the expected magnitude of PAND when the intervention has no effect on the outcome, the top row of the figure indicates that this is not the case.

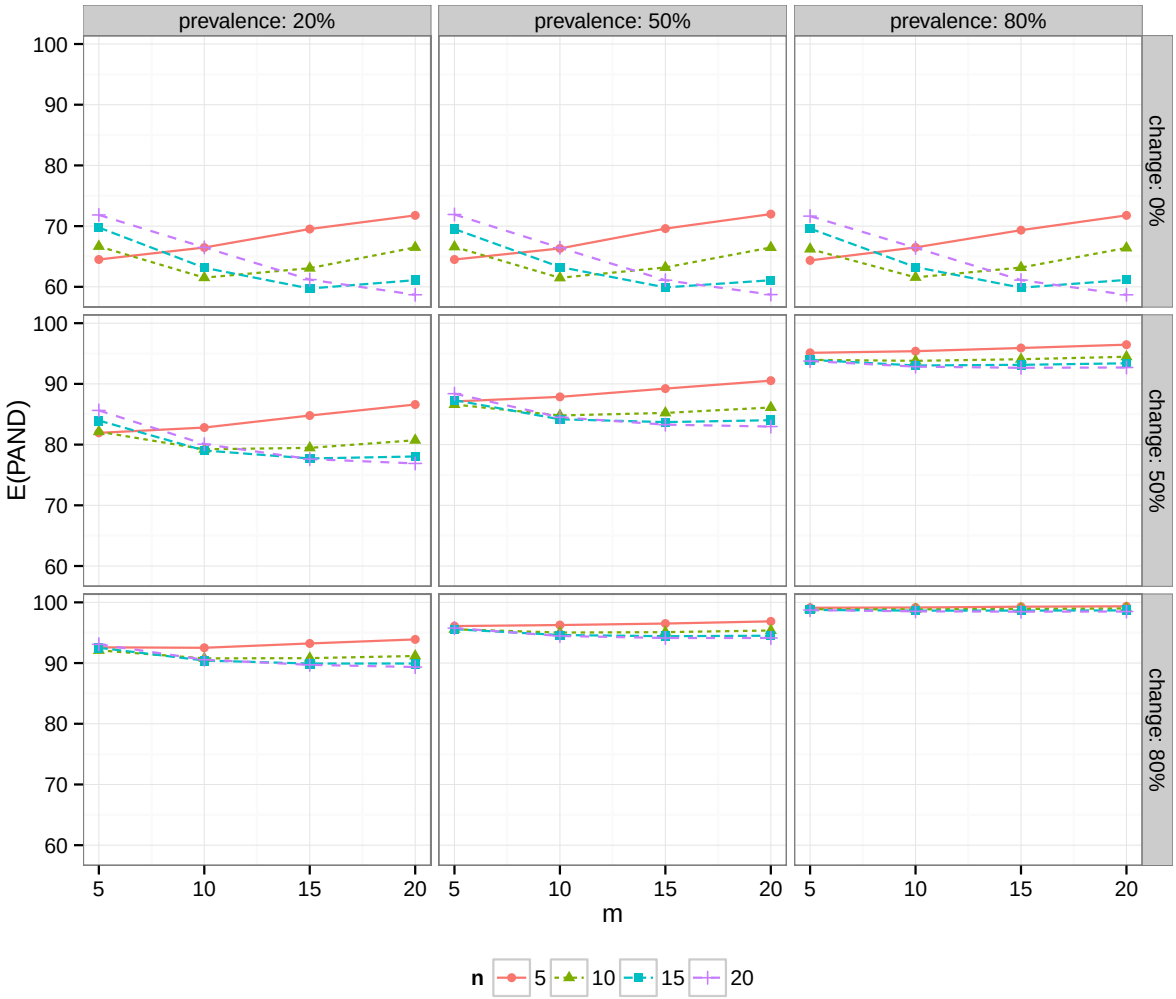

Figure 4. Expected magnitude of PAND based on continuous recording data with $L = 5$, when incidence is once per minute, for varying baseline (m) and treatment phase (n) lengths.

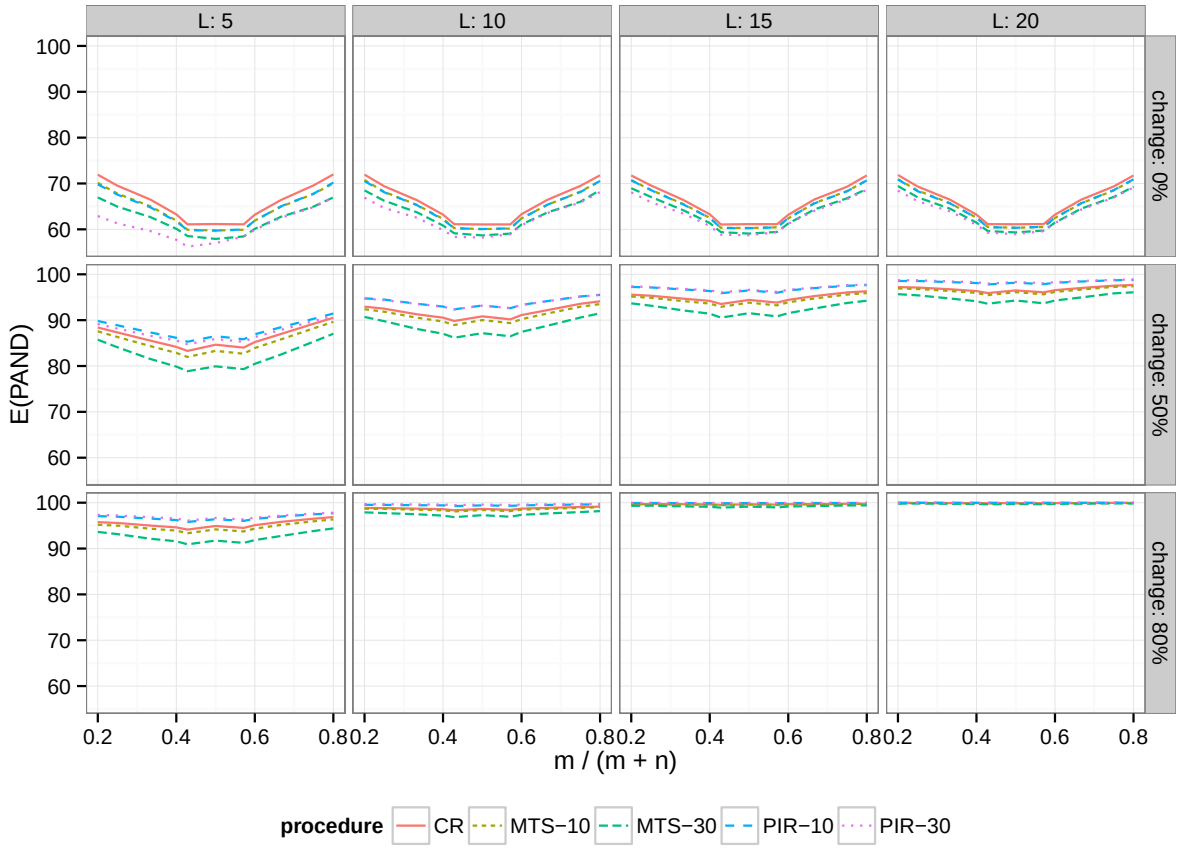

Figure 5. Expected magnitude of PAND based on continuous recording data when prevalence is 50% and incidence is once per minute, for varying session lengths (L) and recording procedures.

Instead, the expected magnitude varies between 59 and 72 when the intervention has no effect. The middle and bottom rows of the figure indicate that PAND becomes less sensitive to sample size when the treatment produces larger effects, though this appears to be mainly because it approaches the maximum level of 100%. The variation in the expected magnitude of PAND appears to be most closely related to the ratio of baseline sessions to the total number of sessions (i.e., $m/(m+n)$); further analysis therefore collapsed across levels of m and n for each unique value of this ratio.

Figure 5 illustrates the sensitivity of PAND to variation in session length and recording procedure, based on the subset of results where prevalence is 50% and

incidence is once per minute. It can be seen that the degree of sensitivity depends on the magnitude of the change from baseline to intervention phase. When the intervention has no effect, PAND is slightly sensitive to what recording procedure is used (continuous recording produces slightly larger values of PAND than MTS with 10 s intervals), but is not affected by the length of the observation session. In contrast, when the intervention reduces the prevalence of the behavior by 50%, the expected magnitude of PAND is quite sensitive to the length of the observation session, with longer sessions leading to higher values for PAND. Finally, when the intervention leads to an 80% reduction in the prevalence of the behavior, the expected magnitude of PAND approaches the ceiling level of 100% regardless of the session length. Taken together, the simulation results demonstrate that PAND is very sensitive to the number of observations in the baseline and treatment phases (particularly when the intervention has no effect), moderately sensitive to observation session length, and slightly sensitive to recording procedure.

IRD

IRD is a linear re-scaling of PAND, and so the behavior of the two measures is generally quite similar. Like PAND, the expected magnitude of IRD depends on the number of observations in each phase. Figure 6 illustrates these relationships for various m and n and varying magnitudes of change between phases; for purposes of clarity, the results are depicted for continuous recording with a session length of five minutes, with prevalence of 50% and incidence of once per minute. When the intervention has no effect, the expected magnitude of IRD is always larger than zero and ranges from 0.12 to 0.29. In contrast to PAND, IRD tends to be larger when the number of observations in the baseline phase is equal to the number in the treatment phase. For larger degrees of change between phases, IRD becomes somewhat less sensitive to the number of observations in each phase.

The degree to which the magnitude of IRD is influenced by the observation session length and the recording procedure closely parallels the results for PAND; results for

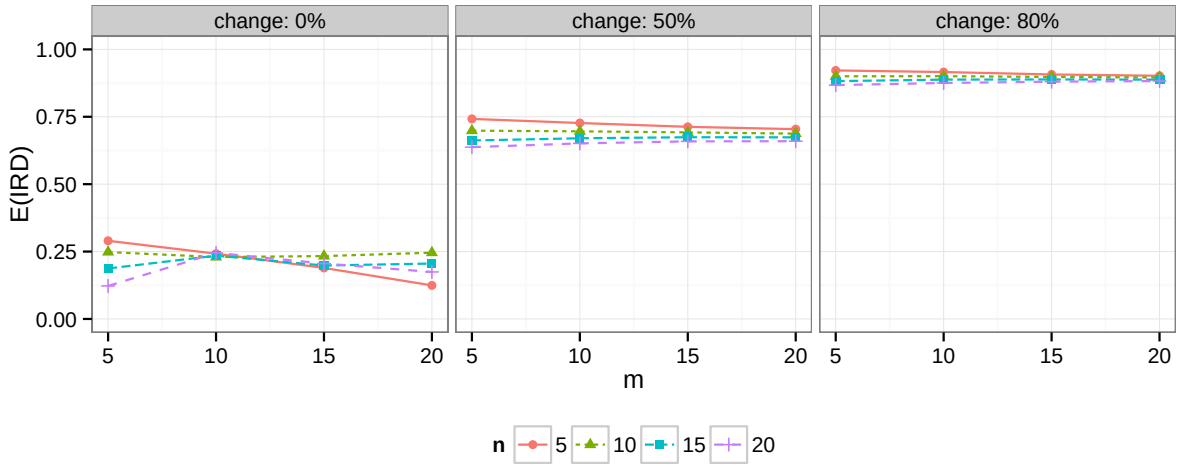

Figure 6. Expected magnitude of IRD based on continuous recording data with $L = 5$, when prevalence is 50% and incidence is once per minute, for varying baseline (m) and treatment phase (n) lengths.

these relationships are therefore omitted. In summary, the expected magnitude of IRD is sensitive to the number of observations in the baseline and treatment phases, moderately sensitive to observation session length, and slightly sensitive to recording procedure.

PEM

The expected magnitude of PEM does not depend on the number of observations in either phase.Furthermore, if the intervention has no effect then the expected magnitude of PEM is always exactly 50%, regardless of the length of the observation sessions or of the recording procedure used to collect outcome data. Consequently, these factors can only effect the magnitude of PEM when there is in fact a change in the score distribution between phases.

To illustrate the sensitivity of PEM to variation in session length and recording procedure, Figure 7 plots the expected magnitude of PEM when the intervention leads to a 50% decrease in behavior, for varying session lengths and recording procedures; each panel displays results for a different combination of prevalence and incidence during

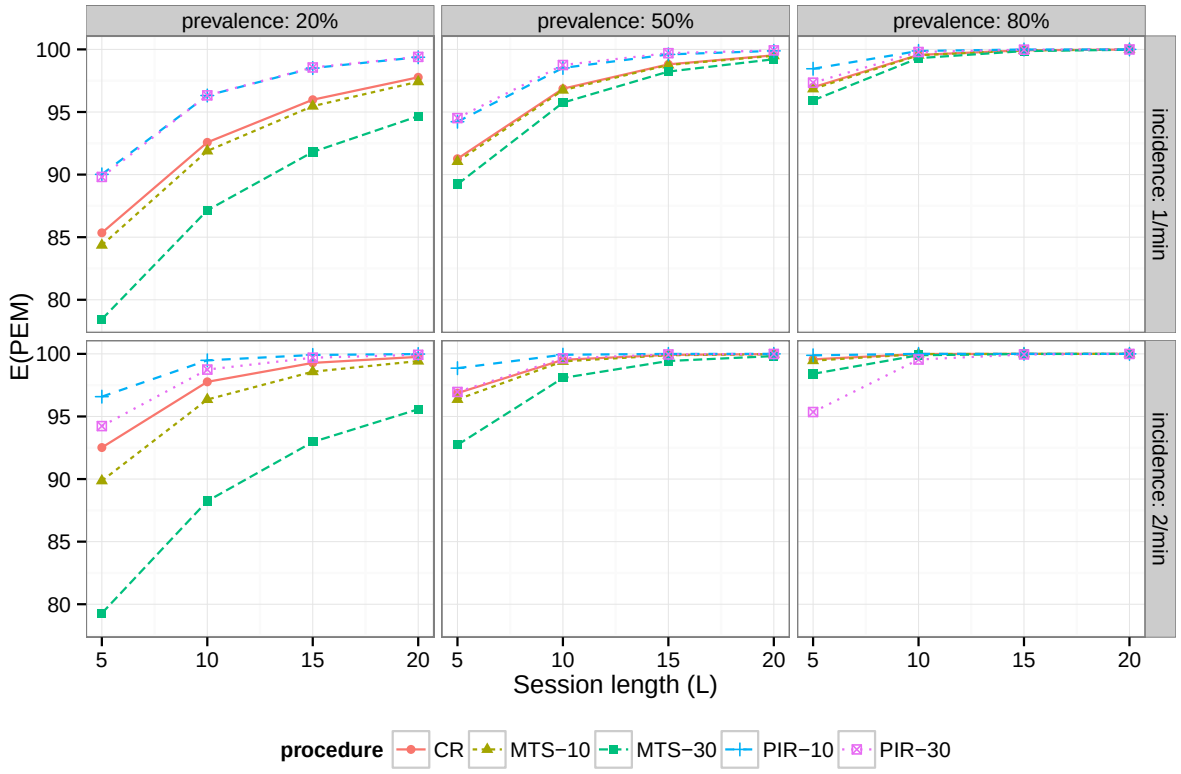

Figure 7. Expected magnitude of PEM.

baseline.[6] It can be seen that, for some types of behavior, the magnitude of PEM is highly sensitive to the length of the observation session and to which recording procedure is used. For instance, for a behavior with baseline prevalence of 20% and baseline incidence of twice per minute, measuring the behavior using 30 s MTS for 5 minute sessions would lead to an expected magnitude of 79, whereas measuring the same behavior using continuous recording for 15 minute sessions would lead to an expected magnitude of 99.

The extent to which PEM is sensitive to these procedural factors depends on the characteristics of the behavior. Specifically, PEM is less sensitive to session length and recording procedure when the behavior has higher levels of baseline prevalence or

[6]When the intervention leads to an 80% decrease in behavior, the expected magnitude of PEM is at or near the ceiling level of 100% across all conditions in the simulation.

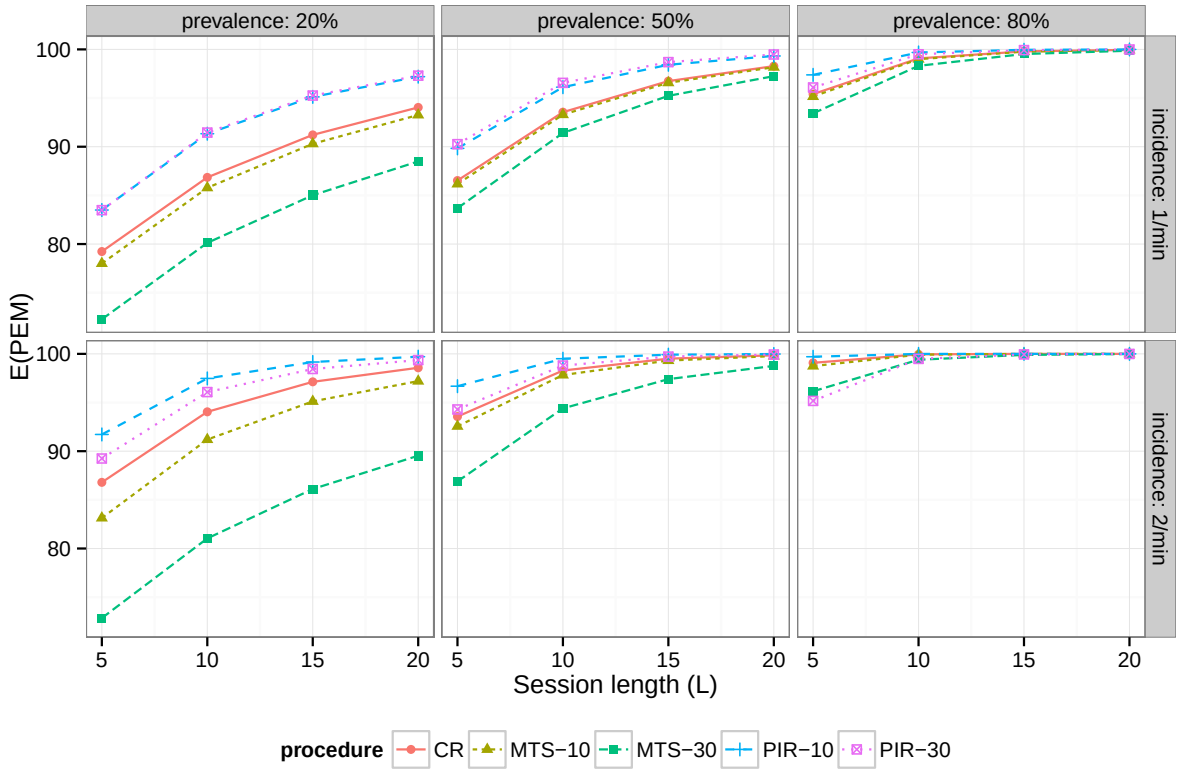

Figure 8. Expected magnitude of NAP.

baseline incidence. However, this reduced sensitivity appears to be largely due to the fact that PEM is at or near the ceiling level of 100% for all session lengths and recording procedures. Thus, it appears that for changes in behavior in the range to which PEM is sensitive, the expected magnitude of this statistic is also influenced by the researcher's choice of observation session length and recording procedure.

NAP

The behavior of NAP is very similar to that of PEM. Like PEM, the expected magnitude of NAP is not influenced by the number of observations in the baseline phase or by the number in the treatment phase. Also like PEM, the expected magnitude of NAP is exactly equal to 50% when the intervention has no effect on the outcome. Figure 8 plots the expected magnitude of NAP when the intervention leads to a 50% decrease in behavior, for varying session lengths and recording procedures. It can be seen that NAP

is highly sensitive to observation session length and recording procedure for certain combinations of behavioral characteristics (particularly for behaviors with lower prevalence), though not for combinations where it is near the ceiling level of 100%. Comparing Figure 8 to Figure 7, it appears that NAP is even more sensitive than PEM to these procedural factors. For instance, when baseline prevalence is 20% and baseline incidence is twice per minute, 5 minute observation sessions using 30 s MTS leads to an expected magnitude of 73, whereas 15 minute sessions using continuous recording leads to an expected magnitude of 97. In summary, it appears that the expected magnitude of NAP is sensitive to observation session length and to recording procedure, so long as the behavioral characteristics and the magnitude of change in behavior are within a range to which NAP is sensitive.

Discussion

Non-overlap measures are a family of statistics that have been proposed for use as effect sizes in single-case designs. Using computer simulations based on a realistic model for systematic behavioral observation data, I have examined the extent to which the magnitude of these statistics is influenced by procedural aspects of a study design. Simulation results demonstrated that PND, PAND, and IRD are all affected by the number of observations in the baseline phase, and the latter two measures are also affected by the number of observations in the treatment phase. All three measures are also sensitive to the length of observation sessions and to the recording procedures used to collect outcome data. T wo other non-overlap measures, PEM and NAP, are unaffected by the number of observations in the baseline or treatment phases, but are sensitive to observation session length and to the choice of recording procedures. Thus, the magnitude of the non-overlap measures is a function partly of arbitrary operational details, chosen by the researcher on the basis of resource availability and feasibility, rather solely of the magnitude of change produced by an intervention. This operational sensitivity makes the non-overlap measures unsuitable for use as effect sizes, because they

do not provide a fair basis for comparison across studies that use different procedures.

The problems raised in this paper add to the growing body of criticism of the non-overlap measures. Researchers have criticized these measures because they lack valid methods to quantify their sampling uncertainty (Shadish et al., 2008), which makes it difficult to apply meta-analytic techniques for synthesis. Others have criticized the non-overlap measures because they do not align well with visual inspection of study results (Wolery et al., 2010).

In light of the operational sensitivity of the non-overlap measures as well as the other criticisms of these measures, I recommend that efforts to compare or synthesize evidence from single-case designs should eschew the use of non-overlap measures as effect sizes. Instead, synthesis efforts should focus on effect size measures that are relatively unaffected by procedural factors that are likely to vary across a collection of SCDs. Extant syntheses of SCDs that make use of non-overlap measures should also be re-examined to determine whether their findings are altered by the use of effect sizes that more clearly quantify the magnitude of treatment effects.

One such measure is the log-response ratio, a well-known effect size measure used in many areas of meta-analysis (Hedges, Gurevitch, & Curtis, 1999). The log-response ratio measures change in proportionate terms, and is appropriate for use with outcomes that are measured on a ratio scale. Systematic direct observation procedures such as continuous recording, momentary time sampling, and event counting do produce such data. Under certain circumstances the log-response ratio is comparable across studies that use different recording procedures (Pustejovsky, 2014b), although studies that use partial interval recording systems present some analytic complications (Pustejovsky & Swan, 2014).

One further implication of this study is that systematic reviews of SCDs should pay more attention to the outcome measurement procedures and study designs on which their findings are based. In particular, researchers should report details regarding the

distribution of observation session lengths, recording procedures, and phase lengths used in the studies included in a systematic review.In addition to simply reporting descriptive information about the range of procedures used, it would also be useful to investigate whether differences in outcome measurement procedures moderate the magnitude of effect sizes in syntheses of SCDs. The simulation results presented in this paper suggest that such moderating effects would occur if the synthesis is based on a non-overlap measure of effect size. Even if another effect size metric such as the log-response ratio is used for synthesis, it would nonetheless be prudent to investigate the potential moderating effects of study procedures, as a test of whether the theoretical properties of the effect size are born out in practice.

There is a long tradition of using non-overlap measures—in particular, PND—to characterize the results of SCDs (cf. Scruggs & Mastropieri, 2012), which has continued despite several stringent critiques that have been leveled (Allison & Gorman, 1994; White, 1987, e.g.). Admittedly, some researchers and journals might still prefer that non-overlap measures be reported as part of primary studies or systematic reviews, because their operational sensitivity does not necessarily negate their utility as descriptive statistics.Indeed, in many respects the non-overlap measures have more in common with hypothesis testing procedures, so it may be more reasonable to interpret these statistics as measures of "strength of evidence" regarding a treatment effect, which account for the sample size and reliability of the outcome measure. However, if non-overlap measures are still to be reported, the results of this analysis should serve as a caution: non-overlap measures are sensitive—sometimes highly so—to operational variation in study design and outcome measurement procedures. Consequently, they should not be interpretted as measures of the magnitude of a treatment effect.

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).

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

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 129–141. doi: 10.1080/17489530802446302

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of consulting and clinical psychology*, *66*(1), 7–18.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual review of psychology*, *52*, 685–716. doi: 10.1146/annurev.psych.52.1.685

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

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

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

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.

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 53–90). Washington, DC: American Psychological Association.

Kelly, M. (1977). A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, *10*(1), 97–101.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. doi: 10.1177/0741932512452794

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly*, *17*(4), 341–389. doi: 10.1521/scpq.17.4.341.20872

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). 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., 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

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

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*(2), 137–148.

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.

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.

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

Pustejovsky, J. E. (2014a). *ARPobservation: Simulating recording procedures for direct observation of behavior.* Retrieved from `http://cran.r-project.org/web/packages/ARPobservation`

Pustejovsky, J. E. (2014b). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological methods*, (In press). doi: 10.1037/met0000019

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. (2014). *Four methods for analyzing partial interval recording data, with application to single-case research.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.r-project.org/`

Rapp, J. T., Colby-Dirksen, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., Britton, L. N., & Colby, A. M. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions*, *22*, 319–345.

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

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. (2012). 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. *Remedial and Special Education*, *8*(2), 24–43.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

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

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

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