Procedural sensitivities of the non-overlap effect sizes for single-case designs

James E. Pustejovsky

The University of Texas at Austin

November 23, 2015

Author note

James E. Pustejovsky, Ph.D. University of Texas at Austin, Austin, TX, USA.

An earlier version of this paper was presented at the 2015 annual convention of the American Educational Research Association, Chicago, IL. Correspondence should be addressed to James E. Pustejovsky, Department of Educational Psychology; University of Texas at Austin; 1912 Speedway, Stop D5800; Austin, TX 78712-1289. Phone: 512-471-0683. Email: pusto@austin.utexas.edu.

Abstract

Non-overlap measures, such as the percentage of non-overlapping data, improvement rate difference, and non-overlap of all pairs, are the most widely used effect sizes in systematic reviews of single-case designs. Although they are often characterized as non-parametric effect sizes, it is unclear whether it is reasonable to interpret these indices as measures of treatment effect magnitude. This study uses computer simulation to investigate the properties of several non-overlap measures when applied to behavioral data collected by systematic direct observation. Results indicate that the magnitude of the non-overlap measures can be strongly influenced by procedural details of the study's design, including the length of the observation session, recording system used to collect data; and phase length. It is argued that non-overlap measures should not be interpreted as effect sizes when applied to behavioral observation data.

Keywords: single-case research; effect size; behavioral observation

Procedural sensitivities of the non-overlap effect sizes for single-case designs

Single-case designs (SCDs) comprise a large and important part of the research base in many areas of special education. For example, in a comprehensive review of focused intervention practices for children with autism, 89% of the 456 identified studies used SCDs (Wong et al., 2015). In light of the size and breadth of this research base, principled methods for systematically reviewing and synthesizing results of SCDs are needed. Careful systematic reviews and syntheses have the potential to bring greater clarity to areas of research where individual studies are small, heterogeneous, and come to discrepant conclusions. Systematic reviews and syntheses have also become an important tool for establishing evidence-based practices and directing clinicians, educators, and caregivers towards more effective interventions. Regarding the latter goal, several prominent research organizations have recently proposed guidelines for establishing evidence-based practices on the basis of evidence from SCDs, including the Council for Exceptional Children's Division for Research (Cook et al., 2015; Horner et al., 2005), and the What Works Clearinghouse (Kratochwill et al., 2012). 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.

A primary decision that must be made in any quantitative synthesis of intervention research (whether based on between-groups designs or SCDs) is what effect size to use to quantify the magnitude of treatment effects. Effect size indices are the basic unit of analysis in a quantitative synthesis—the metric on which study results are combined and compared. Consequently, it is important for both producers and consumers of systematic reviews to understand their proper interpretation and their limitations.

If an effect size index is to have a valid interpretation in terms of treatment effect magnitude, it must provide a reasonable basis for comparison from one study to another (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges, 2008; Lipsey & Wilson, 2001). To do so, an effect size must be relatively insensitive to incidental features of a study's design, 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. If an effect size is procedurally sensitive, it becomes difficult to make sensible comparisons across studies or to combine the results of several studies because the metric of the comparison is not consistent. Procedural insensitivity is thus a basic and fundamental property for an effect size index to be useful. However, relatively little attention has been paid to whether effect sizes that are commonly used for summarizing the results of SCDs have this property.

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), between-case standardized mean differences (Shadish, Hedges, & Pustejovsky, 2014), and measures based on piece-wise linear regression models (e.g., Center, Skiba, & Casey, 1985; Maggin, Swaminathan, et al., 2011), have been developed from specific, parametric statistical models. Most of these models are premised on the assumption that the dependent variable is normally distributed, but this assumption is often viewed as inappropriate for data from SCDs.

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, 1987b), improvement rate difference (Parker, Vannest, & Brown, 2009), and a number of others. One perceived advantage of these measures is that they are non-parametric, meaning that they are not based upon distributional assumptions about the dependent variable (Parker, Vannest, & Davis, 2011). The NOMs are also viewed as being more intuitively interpretable than parametric effect sizes because they are defined in terms of overlap percentages (i.e., 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).

Despite their popularity and wide application, little previous research has examined the properties of the NOMs as effect size indices. From a statistical point of view, the fact that the NOMs are not developed under specific assumptions about the distribution of the data makes it difficult to judge whether—and under what circumstances—these indices are sensitive to procedural aspects of a study's design. The present study aims to fill this gap, by studying the characteristics of the NOMs using data simulated from a realistic model for direct observation of behavior.

Behavioral measures derived from systematic direct observation are the most common type of dependent variable 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 (Ayres & Gast, 2010). Each of these procedures can be used for shorter or longer periods of observation, and the number of observation sessions will also vary from study to study. Thus, in order to have a fair basis for comparisons between and synthesis of SCDs that use behavioral outcomes, an effect size index should be insensitive to the researcher's decisions about how the dependent variable is measured.

In order to study the properties of the NOMs 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 actual process of observing a behavior stream and recording data as one does so. As a result, it provides a way to emulate the distinctive features of real, empirical behavioral observation data.

Using the alternating renewal process model, this study investigates the extent to which extant NOMs are sensitive to procedural features of a single-case study, including the length of the observation sessions during which outcome measures are collected, the procedure used to record behavioral observations, and the number of sessions during baseline and during treatment. The next section reviews the NOMs that have been proposed as effect size measures for 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.

Non-overlap measures

This section describes several of the main NOMs that have been proposed as effect size measures for SCDs, including 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 Tau. Parker and colleagues (2011) provide a more expansive review of the NOMs, including worked examples of how to calculate each effect size based on graphed data. Extensions of certain measures have been proposed that accommodate time trends in the outcome measures (Manolov & Solanas, 2009; Parker, Vannest, Davis, & Sauber, 2011; Wolery, Busick, Reichow, & Barton, 2010); for sake of simplicity, the present discussion is limited to NOMs that are appropriate for data that do not display time trends. This section also describes available guidelines for characterizing effects as “small,” “medium”, or “large,” which provide one way to judge whether the procedural sensitivities of the NOMs are consequential.

Each of the NOMs is defined in terms of a comparison between a baseline phase and a treatment phase. It will therefore suffice to consider a simple AB-type design, with a single baseline phase and a single treatment phase. Throughout, I assume that the dependent variable is defined in such a way that larger values correspond to more beneficial outcomes, so that increases are desirable. To rule out potential ambiguity in their definitions, the supplementary materials for this article provide exact mathematical formulas for each of the NOMs.

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 exceed the highest measurement from the baseline phase (Scruggs et al., 1987b). 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).

Since it was first proposed, PND has been widely criticized (Shadish, Rindskopf, & Hedges, 2008; White, 1987; Wolery et al., 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—that is, its average value across repeated samples—is strongly influenced by the number of observations in the baseline phase. More specifically, longer baseline phases will tend to result in smaller values of PND, even when treatment has no effect at all. They argued that this dependence on sample size makes the statistic unsuitable for use as an effect size metric. Despite these objections, PND remains by far the most commonly applied effect size in systematic reviews of SCDs (Maggin, O'Keeffe, et al., 2011; Scruggs & Mastropieri, 2013).

Percentage exceeding the median

In response to some of the criticisms of PND, Ma (2006) proposed an alternative that uses the median of the baseline phase (rather than 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. Like PND, PEM ranges in principle from 0 to 100%. Unlike PND, the expected magnitude 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.

Percentage of all non-overlapping data and robust improvement rate difference

Parker, Hagan-Burke, and Vannest (2007) proposed the percentage of all non-overlapping data (PAND) and the closely related phi coefficient. The authors argued that the latter offers the advantage of having a known sampling distribution. In later work, the same authors proposed an effect size called the robust improvement rate difference (RIRD), which is a "robust" version of the phi coefficient where overlapping observations are evenly divided between the lower left and upper right cells of the table (Parker et al., 2009; Parker, Vannest, & Davis, 2011). Later work by the same authors indicated that PAND has been superseded by RIRD (Parker, Vannest, & Davis, 2014), and so I focus only on the latter. The supplementary materials provide further details about PAND.

Unlike PND and PEM, the logical range of the RIRD statistic is not obvious. Furthermore, the simulation study will demonstrate that, when treatment has no effect, the expected magnitude of RIRD depends on the lengths of each phase. Based on a comparison between the RIRD statistics and expert visual assessments, Parker and colleagues (2009) proposed 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 (p. 147).

Non-overlap of all pairs and Tau

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 exceeds the measurement from the baseline phase. Pairs of data points that are exactly tied are counted with a weight of 0.5. The logical range of NAP is from 0 to 100%, with a stable expected magnitude of 50% when treatment has no effect on the outcome. 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).

Parker, Vannest, Davis, and Sauber (Parker, Vannest, Davis, et al., 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, which is equivalent to a linear re-scaling of NAP to a scale of -1 to 1. Tau has an expected magnitude of 0 when treatment has no effect on the outcome. Because NAP and Tau are so closely related, the simulation study focuses on the former measure only.

The alternating renewal process model

In order to study the properties of the NOM effect sizes when applied to behavioral observation data, a statistical model is needed that is flexible enough to capture the relationships between the details of how the behavior is measured and the distribution of the resulting data. Standard probability distributions such as the normal, binomial, or Poisson are not rich enough to model these relationships. The simulations described in the following 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 behavioral episodes that each last some length of time. With such behaviors, the researcher is typically most interested in the behavior’s *prevalence*, or the overall proportion of time that it occurs, and continuous recording or an interval-based system is often used for measurement. Ontask behavior 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*, or overall rate of occurrence per unit of time, and frequency counting is often used for measurement. Problem behaviors such as hitting or 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, which emulates the pattern of behavior that an observer would actually see during the course of a session. Each behavior stream consists of episodes of behavior separated by spans of inter-response time. The duration of each behavioral episode is generated randomly, according to a certain probability distribution; similarly, the length of each inter-response time is generated randomly, according to a different probability distribution. For state behaviors, the average duration of the behavior will typically be long enough to be of consequence, whereas for event behaviors, the average duration of each episode will be close to zero. 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 the simulated behavior stream based on the rules of a specified 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 calculated as the total number of simulated 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, each of a specified length (e.g., 15 s). 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 with behavior. Pustejovsky and Runyon (2014) provide further details about the recording systems and summary measurements.

Simulation design

Using the alternating renewal process model, I conducted a simulation study to examine the extent to which the NOMs 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 are focused on state behaviors, where the behavior's prevalence is the primary characteristic of interest. Another simulation study, focused on event behaviors, is reported in the supplementary materials.

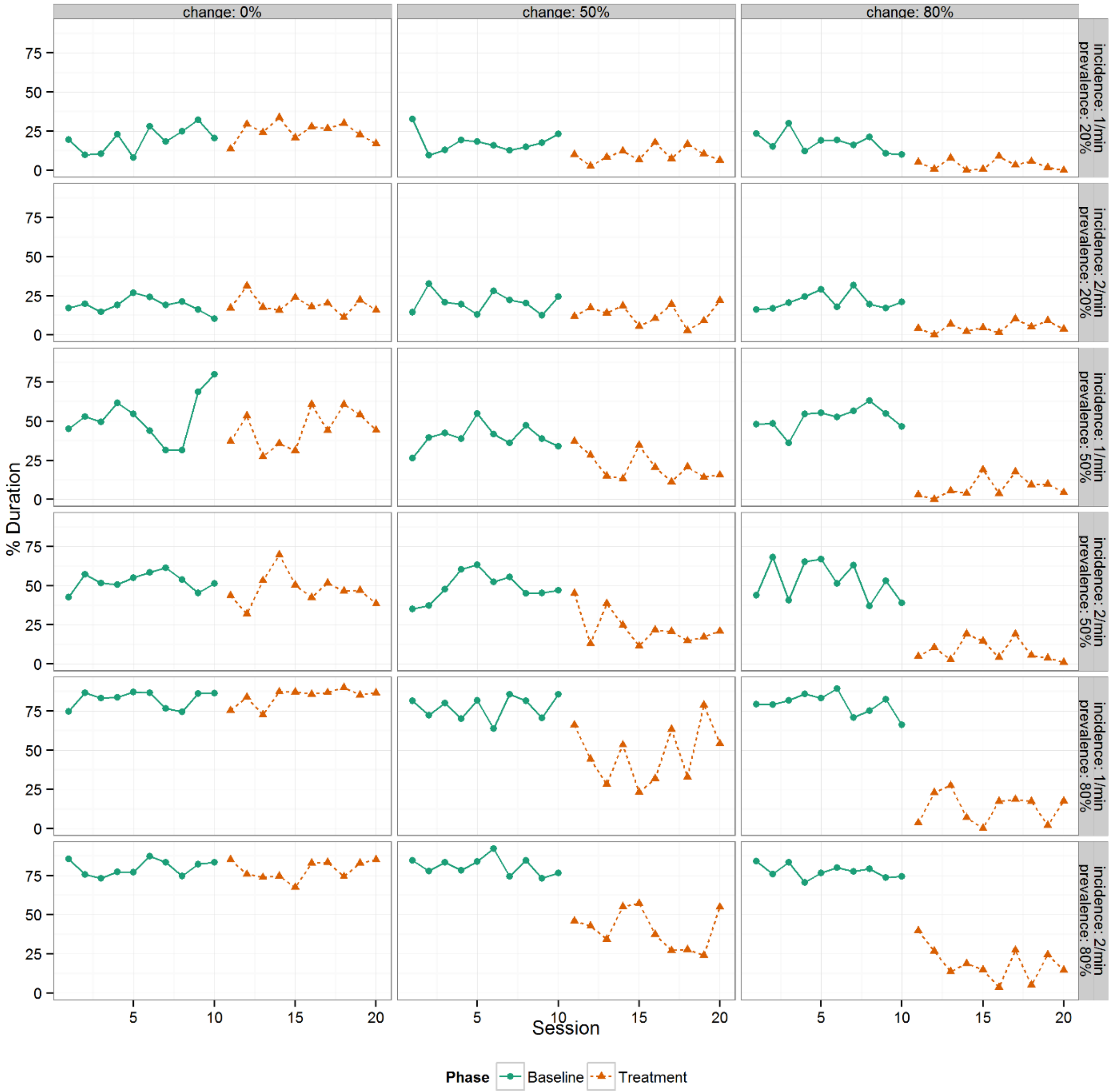
Data-generating model

The simulation used the alternating renewal process model to generate realistic measurements of a state behavior, in the context of a SCD study with a simple AB design. To fully operationalize the alternating renewal process model, several further details had to be specified. Although prevalence is the main behavioral characteristic of a state behavior, its incidence is relevant as well, 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 episode duration and average inter-response time used to generate behavior streams. The properties of the behavior stream also depend to some extent on the parametric form of the distributions used to simulate episode durations and inter-response times.

I chose values for the parameters of the behavior to represent a range of conditions, and then verified the plausibility of the conditions by visual inspection of the simulated data. For sake of simplicity, I assumed that episode durations and inter-response times followed exponential distributions. 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. 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 incidence of the behavior during the baseline phase to once per minute or twice per minute. Many SCDs focus on behaviors in 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 1 displays examples of simulated SCDs for each combination of prevalence, and incidence, and change in behavior. These examples were generated using continuous recording with 10 min observation sessions and 10 sessions in each phase. The reader may judge for themselves whether the simulated data resemble the data from real SCDs.

Figure 1. Simulated SCDs based on the alternating renewal process model, using continuous recording for 10 min sessions, for varying levels of prevalence, incidence, and change in behavior



Procedural factors

The simulation examined the effects of three aspects of the measurement procedures: recording system, 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 closely resemble the procedures used in actual SCDs.

Continuous recording (CR), MTS, and PIR are the main systems for recording direct 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 (Mudford, Taylor, & Martin, 2009; Rapp et al., 2007). 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, a recent synthesis of SCDs examined the effect of functional behavior assessment interventions on student problem behavior (Gage, Lewis, & Stichter, 2012); of the cases included in the synthesis, observation sessions ranged from 5 to 60 min per session and 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 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 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 of baseline and treatment lengths.

Simulation procedures

I conducted the simulation using the ARPobservation package (Pustejovsky, 2014a) for the R statistical computing environment. For each combination of factor levels, I simulated 10,000 AB designs and calculated the PND, PAND, RIRD, PEM, and NAP measures. I then averaged the results across replications in order to estimate the expected magnitude of each measure. The computer code that implements the simulation and full numerical results are available in the supplementary materials.

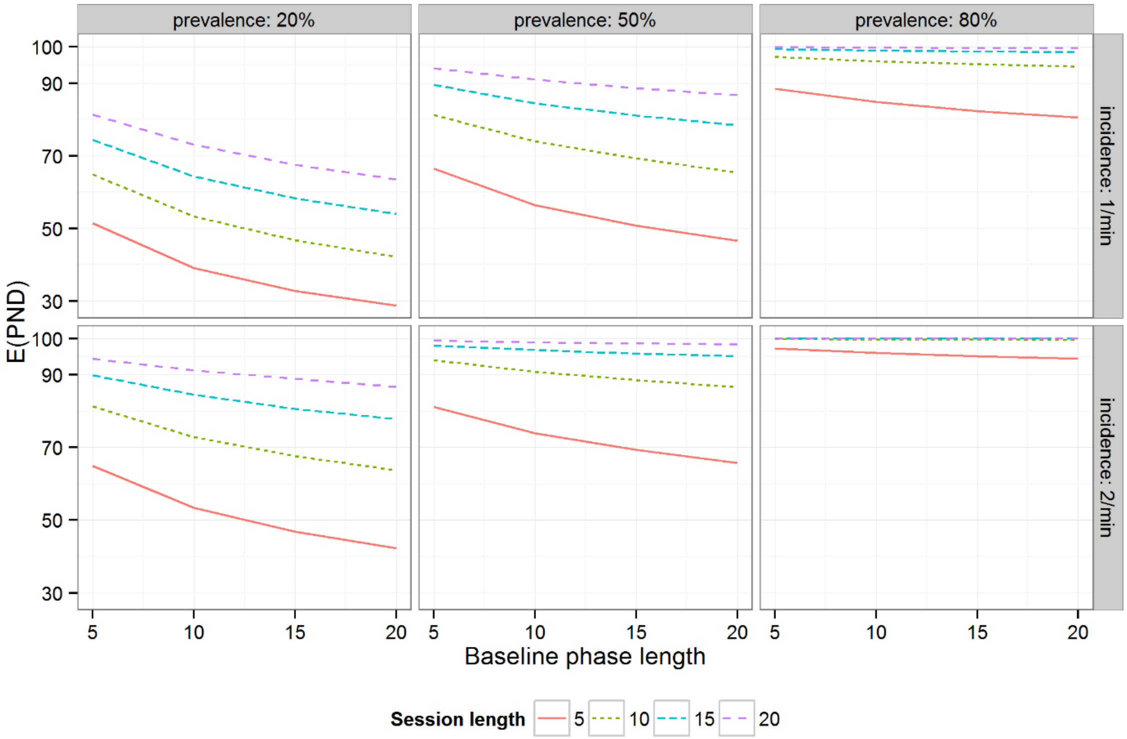
Simulation Results

I present the results of the simulation study in the form of figures that illustrate the degree to which the expected magnitude of each NOM varies as a function of the procedural aspects of the design. For measures whose expected magnitude is known to be unaffected by phase length, results are averaged across the levels of the baseline and treatment phase lengths. For clarity of presentation, some of the figures present results for selected subsets of the conditions; in these cases, the results that are presented are generally consistent with the other simulation conditions, except when otherwise noted.

Percentage of non-overlapping data

Allison and Gorman (1994) demonstrated that, when treatment has no effect (i.e., change of 0%), the magnitude of PND depends strongly on the number of observations in the baseline phase. The results of the present simulation replicated this finding. For continuous recording, the expected magnitude of PND was exactly equal to 100% / (m + 1), where m denotes the length of the baseline phase, regardless of the length of the session or the characteristics of the behavior. The expectations based on other recording systems differed slightly due to the possibility of exact ties, which do not occur with continuous recording data.

Figure 2. Expected magnitude of PND based on continuous recording data, when treatment leads to a 50% change, for varying session lengths and baseline phase lengths



When treatment produces beneficial effects, PND remains sensitive to baseline length and recording system and also becomes sensitive to length of the observation session. Figure 2 illustrates the degree of sensitivity, displaying the expected magnitude 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, 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 94%, depending on the session length and length of the baseline phase. By the guidelines of Scruggs and Mastropieri (1998), an intervention might therefore appear to be “ineffective” or “very effective” depending solely on features of the study’s design.

Percentage of all non-overlapping data and robust improvement rate difference

The properties of PAND and RIRD are quite similar. Parker and colleagues (Parker et al., 2014) argued that PAND has been superseded by RIRD, and so this section presents results only for the latter measure. Results for PAND can be found in the supplementary materials.

Figure 3. Expected magnitude of RIRD based on continuous recording data with 5 min observation sessions, when incidence is 1/min, for varying baseline and treatment phase lengths

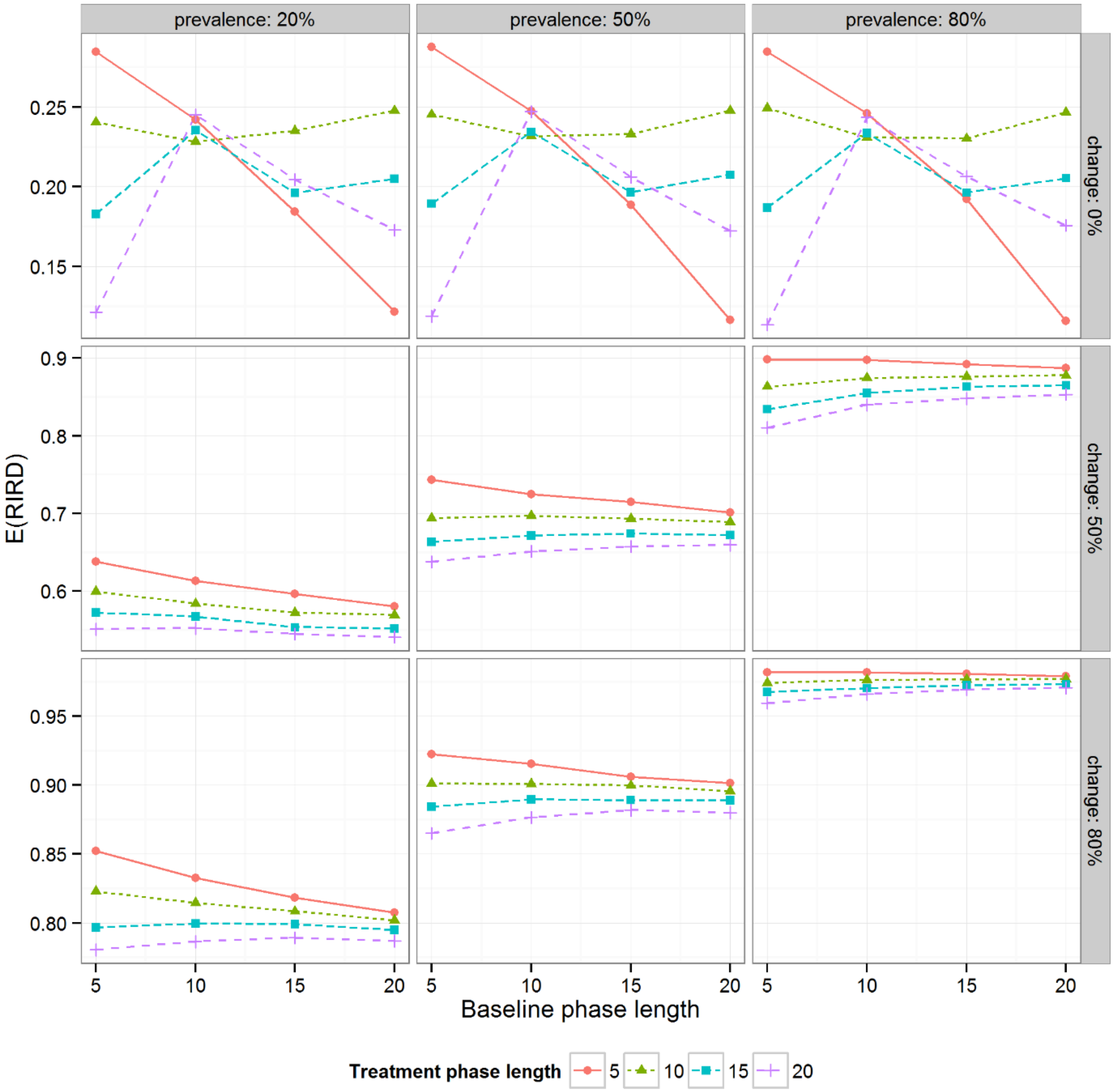


Figure 3 depicts the expected magnitude of RIRD as a function of the number of observations in the baseline phase and in the treatment phase, for the subset of results where continuous recording is used for 5 min sessions and where incidence is once per minute. The top row of the figure indicates that the expected magnitude of RIRD varies between .11 and .29 when the treatment has no effect. For larger degrees of change between phases, RIRD becomes somewhat less sensitive to the number of observations in each phase; for instance, when treatment produces a 50% change in behavior, RIRD ranges from .64 to .74.

Figure 4. Expected magnitude of RIRD based on continuous recording data with 10 sessions in the baseline phase, when prevalence is 50% and incidence is 1/min, for varying session lengths and recording systems

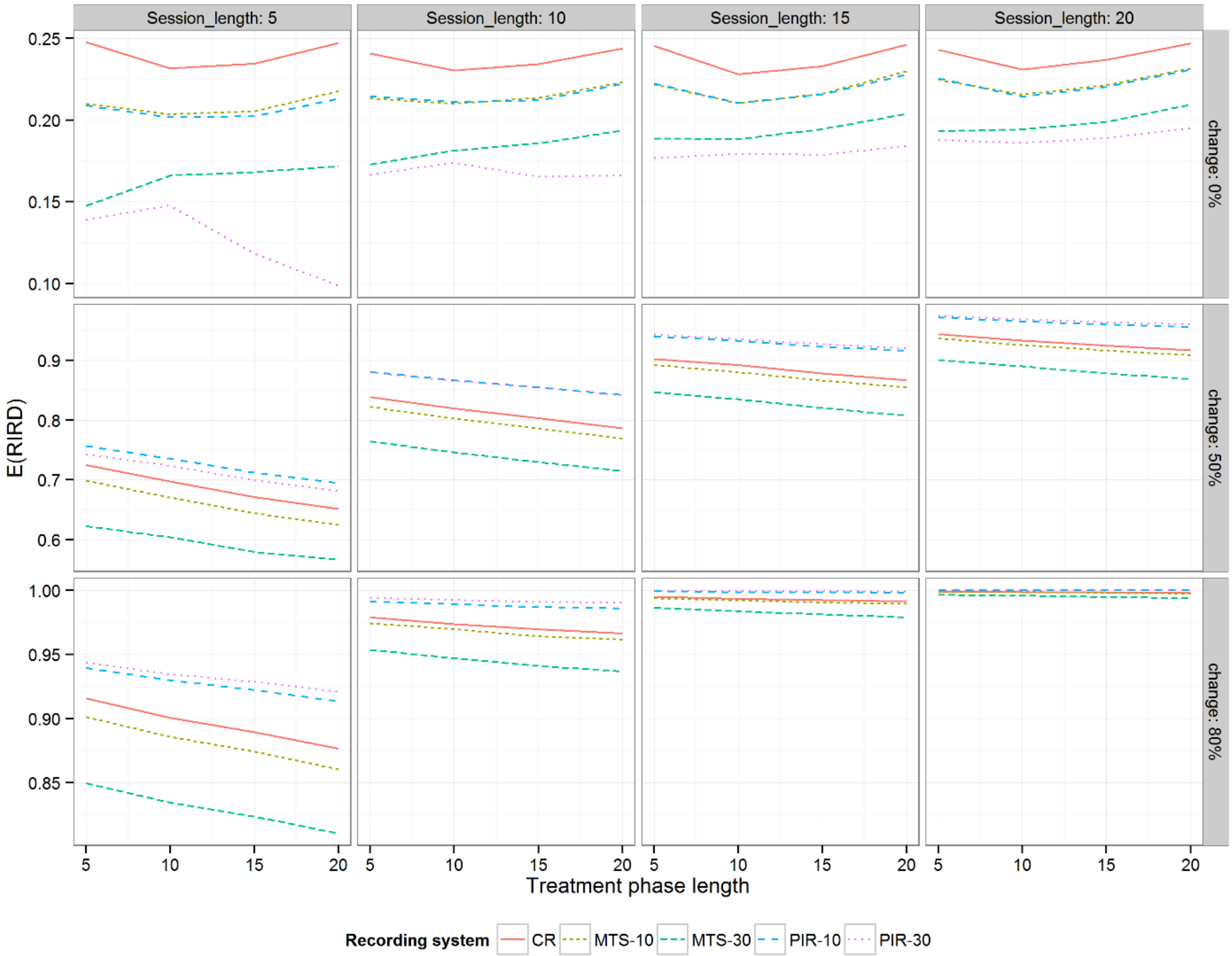


Figure 4 illustrates the sensitivity of RIRD to variation in session length and recording system, based on the subset of results where prevalence is 50%, incidence is once per minute, and the baseline phase includes 10 sessions. It can be seen that the degree of sensitivity depends on the magnitude of the change from baseline to intervention phase. When treatment has no effect, RIRD is sensitive to what recording system is used but is only slightly affected by the length of the observation session. In contrast, when the treatment reduces the behavior by 50%, the expected magnitude of RIRD is quite sensitive to the length of the observation session, with longer sessions leading to values that are more likely to indicate “large” effects; RIRD also remains sensitive to the recording system. Finally, when treatment leads to an 80% reduction in the prevalence of the behavior, the expected magnitude of RIRD approaches the ceiling level of 100% regardless of the session length or recording system.

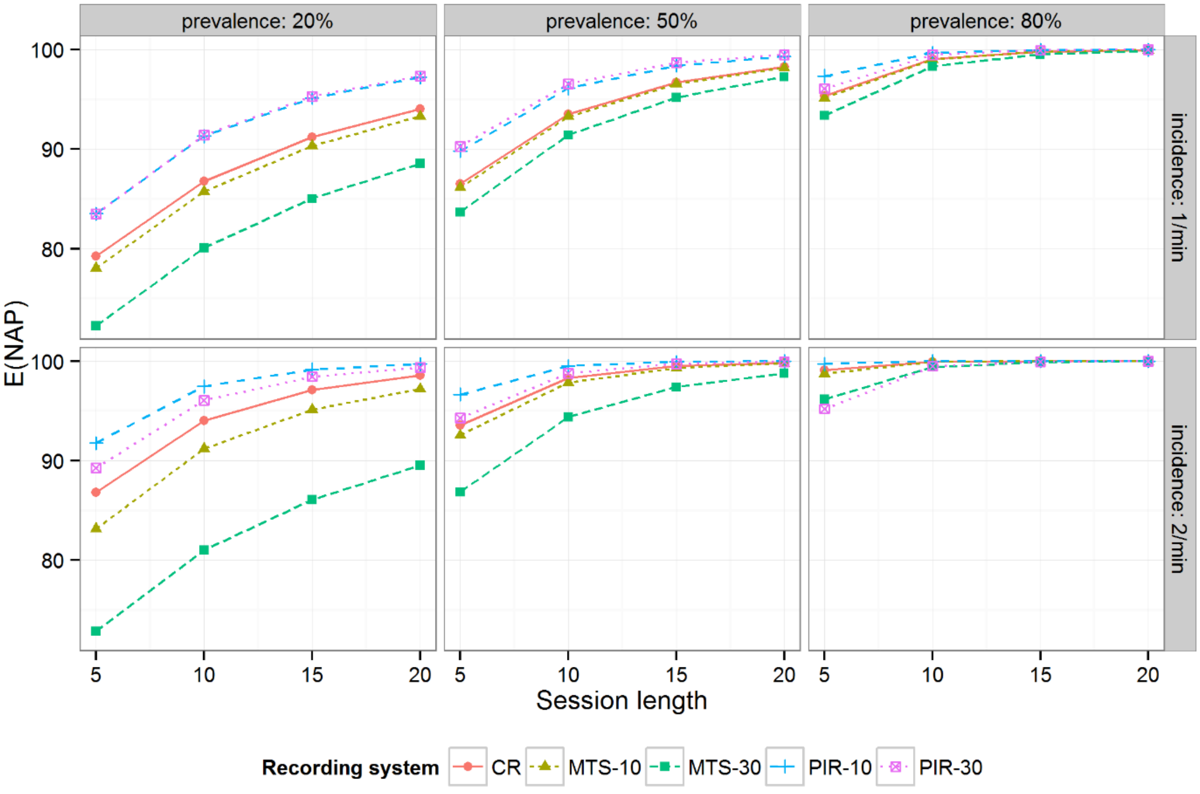
Non-overlap of all pairs and percentage exceeding the median

The expected magnitude of NAP is unaffected by the number of observations in the baseline phase or the treatment phase. Furthermore, if treatment has no effect on the behavior then the expected magnitude of NAP is always exactly 50%, regardless of the length of the observation sessions or of the recording system used to collect outcome data. Consequently, these factors are only matter when there is a change in the score distribution between phases.

To illustrate the sensitivity of NAP to variation in session length and recording system, Figure 5 plots its expected magnitude when treatment leads to a 50% decrease in behavior, for varying session lengths and recording systems; each panel displays results for a different combination of prevalence and incidence during baseline. (When treatment leads to an 80% decrease in behavior, the expected magnitude of NAP is at or near the ceiling level of 100% across all conditions in the simulation.) 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 sessions are 10 min, using 30 s MTS leads to an expected magnitude of 81% (a “medium” effect), whereas using 10 s PIR leads to an expected magnitude of 96% (a “large” effect).

The extent to which NAP is sensitive to these procedural factors depends on the characteristics of the behavior. NAP is less sensitive to session length and recording procedure when the behavior has higher levels of baseline prevalence or baseline incidence. However, this reduced sensitivity appears to be 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

Figure 5. Expected magnitude of NAP for various recording systems and session lengths, when treatment leads to a 50% change in behavior



which it is sensitive, the expected magnitude of NAP is sensitive to the choice of observation session length and recording system.

The results for the PEM statistic are very similar to those for NAP. Like NAP, PEM is unaffected by the number of observations in either phase and its expected magnitude is 50% when treatment has no effect on the outcome. PEM is also sensitive to the choice of observation session length and recording procedure when the behavioral characteristics are within the range where PEM is at all sensitive. The supplementary materials provide further details about the results for PEM.

Discussion

Using computer simulations based on a realistic model for systematic behavioral observation data collected in a SCD, I have examined the extent to which NOMs are influenced by operational procedures. Simulation results demonstrated that the magnitude of these measures is a function partly of arbitrary procedural details—likely selected by the researcher on the basis of resource availability and feasibility—rather than solely of the magnitude of change produced by an intervention. The procedural sensitivities of the NOMs make them unsuitable for use as effect sizes for quantifying treatment effects on behavioral outcomes, because they do not provide a fair basis for comparison across studies that use different procedures.

The problems with the NOMs identified in this paper add to the 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, Karich, & Symons, 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.

Although the operational sensitivities of the NOMs makes them inappropriate for use as effect sizes, this does not necessarily negate their utility as descriptive statistics. Results of the simulation study indicated that the NOMs behave in some respects like test statistics (e.g., the two-sample t-statistic), which are a function of both effect magnitude and the precision with which that magnitude can be estimated. Thus, it may be more reasonable to interpret them as measures of *strength of evidence* that the treatment produced a non-zero effect—a distinctly different concept than treatment effect magnitude. In fact, the developers of PND have made similar arguments, describing PND as “a measure of the *tangibility* (or ‘convincingness’) of the effect” (Scruggs, Mastropieri, & Casto, 1987a, p. 41), while still being related to treatment effect magnitude (Scruggs & Mastropieri, 2013). This interpretation is also consistent with the concern raised by Parker and colleagues over the statistical power of the NOMs (Parker, Vannest, & Davis, 2011) and with the premise that NOMs should be in agreement with visual inspection of single-case data (e.g., Parker & Vannest, 2012).

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: the NOMs may be less (or more) sensitive to operational features of the study design for patterns of behavior and measurement procedures outside of those examined here. More fundamentally, the findings hinge on the extent to which the alternating renewal process model is a reasonable 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 currently available to investigate the model's distributional assumptions. Until such evidence can be collected, the credibility of the model rests on its face validity, as assessed for example by visual inspection of the simulated data from Figure 1.

Given that this study used simulation methods, a crucial avenue of further research is to examine the properties of the NOMs using real empirical data. An investigation of the associations between the magnitude of the NOMs and the operational characteristics of a set of real SCDs would provide an important and independent source of evidence regarding the problems raised in the present study.

Implications for applied research

In light of the operational sensitivity of the NOMs as well as other criticisms that have been raised about these measures, efforts to compare or synthesize evidence from SCDs with behavioral outcomes should not use non-overlap measures as effect sizes. Instead, synthesis efforts should focus on effect size metrics that are relatively unaffected by study procedures that are likely to vary across a collection of SCDs. Existing syntheses of SCDs that use NOMs should 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 other research domains (e.g., Hedges, Gurevitch, & Curtis, 1999). The log-response ratio measures change in proportionate terms and is appropriate for use with ratio scale outcomes, such as direct observation data based on continuous recording or frequency counting. The log-response ratio has a close relationship with percentage change measures of effect size (Campbell & Herzinger, 2010), and thus may be intuitively appealing to behavioral researchers and clinicians who conceptualize treatment effects in such terms. Additionally, under certain circumstances the log-response ratio is comparable across studies that use different recording systems (Pustejovsky, 2014b), although studies that use partial interval recording systems present certain complications (Pustejovsky & Swan, 2015).

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 phase lengths used 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.

Implications for methodological research

Despite the problems with existing NOMs, some methodological researchers may be nonetheless be interested in further developing of overlap-based measures of effect size for SCDs. This study illustrates the importance of validating any such new developments using plausible data-generating models such as the alternating renewal process. Furthermore, better statistical models and a stronger understanding of the psychometric properties of other types of outcome data used in single-case research, such as academic performance measures, would contribute to improved methods of synthesizing SCDs.

Conclusion

There is a long tradition of using non-overlap measures—and especially PND—to characterize the results of SCDs, which has continued despite stringent critiques (e.g., Allison & Gorman, 1994; Shadish et al., 2008; White, 1987). Given this history and the familiarity of non-overlap methods among applied researchers, it seems likely that researchers conducting syntheses of SCDs might continue to prefer to report NOMs as part of primary studies or systematic reviews of single-case research. The What Works Clearinghouse standards for SCDs also recommends the use of multiple effect size measures, which may also encourage continued use of NOMs (Kratochwill et al., 2012). If they do continue 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 behavioral outcome measurement procedures. Consequently, they should not be interpreted as measures of treatment effect magnitude.

References

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., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9780470743386

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., & 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). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387–400. doi:10.1177/002246698501900404

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC’s standards for classifying the evidence base of practices in Special Education. *Remedial and Special Education*, *36*(4), 220–234. doi:10.1177/0741932514557271

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.

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. doi:10.1890/0012-9658(1999)080%5B1150:TMAORR%5D2.0.CO;2

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

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

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., 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

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

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. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education*, *21*(3), 254–265. doi:10.1007/s10864-012-9153-1

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. (2014a). ARPobservation: Simulating recording procedures for direct observation of behavior. R package Version 1.0. 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*, Advance online publication. 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. (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

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: a re-evaluation. *Behavioral Interventions*, *22*(4), 319–345. doi:10.1002/bin.239

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. (1987a). Reply to Owen White. *Remedial and Special Education*, *8*(2), 40–42. doi:10.1177/074193258700800208

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987b). 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., 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

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, 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

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