An examination of measurement procedures and characteristics of baseline outcome data in single-case research

James E. Pustejovsky[1], Daniel M. Swan[1], & Kyle W. English[1]

[1] University of Texas at Austin

August 2, 2019. Forthcoming in *Behavior Modification*. This paper is not the version of record and may not exactly replicate the final, published version of the article. The final article will be available, upon publication, via its DOI: 10.1177/0145445519864264

Author Note

James E. Pustejovsky, Daniel M. Swan, and Kyle W. English, Educational Psychology Department, University of Texas at Austin.

This work was supported by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education. The authors thank Erin Barton, Evan Dart, Nicholas Gage, Jennifer Ledford, Daniel Maggin, and Karrie Shogren for graciously sharing data from their systematic reviews. Complete raw data and R code for replicating this paper are available at https://osf.io/n9jud/

Correspondence concerning this article should be addressed to James E. Pustejovsky, 1912 Speedway, MS D5800, Austin, TX 78712. E-mail: pusto@austin.utexas.edu

Abstract

There has been growing interest in using statistical methods to analyze data and estimate effect size indices from studies that use single-case designs (SCDs), as a complement to traditional visual inspection methods. The validity of a statistical method rests on whether its assumptions are plausible representations of the process by which the data were collected, yet there is evidence that some assumptions—particularly regarding normality of error distributions—may be inappropriate for single-case data. To develop more appropriate modelling assumptions and statistical methods, researchers must attend to the features of real SCD data. In this study, we examine several features of SCDs with behavioral outcome measures in order to inform development of statistical methods. Drawing on a corpus of over 300 studies, including approximately 1800 cases, from seven systematic reviews that cover a range of interventions and outcome constructs, we report the distribution of study designs, distribution of outcome measurement procedures, and features of baseline outcome data distributions for the most common types of measurements used in single-case research. We discuss implications for the development of more realistic assumptions regarding outcome distributions in SCD studies, as well as the design of Monte Carlo simulation studies evaluating the performance of statistical analysis techniques for SCED data.

Keywords: single-case research; behavioral observation; systematic review; alternating renewal process

An examination of measurement procedures and characteristics of baseline outcome data in single-case research

Single-case designs (SCDs) serve as a cornerstone of research on behavior modification and also play an important role in certain areas of special education (Gast & Ledford, 2018), clinical and school psychology (Kazdin, 2011), and communication sciences (Byiers, Reichle, & Symons, 2012), among other fields. To draw conclusions in studies that use SCDs, researchers have traditionally relied upon visual inspection of graphed data (Horner & Swoboda, 2014; Smith, 2012). However, there has also long been interest in using statistical methods to analyze data from SCDs, and a large and growing array of statistical approaches are now available (for a recent review, see Manolov & Moeyaert, 2017).

Development of statistical methods for SCD data has been bolstered as researchers have increasingly sought to use systematic reviews and research syntheses of SCDs to inform evidence-based practice (e.g., Horner et al., 2005; Kratochwill & Stoiber, 2002; Wong et al., 2015). Researchers have developed methods for quantifying the magnitude of treatment effects, in the form of effect size indices and meta-analytic models, in order to synthesize findings from collections of SCD studies (Pustejovsky & Ferron, 2017). Production of systematic reviews and meta-analyses of SCD studies has also increased markedly over the past two decades (Jamshidi et al., 2018; Maggin, O'Keeffe, & Johnson, 2011). Thus, there is a need to scrutinize the statistical analysis methods used in such reviews, as well as methods used to analyze data from individual SCDs.

The validity of a statistical method hinges on whether its assumptions are plausible representations—or at least reasonable approximations—of the process by which the data were collected. Many well-known statistical methods for SCD data, such as the effect size methods proposed by Center, Skiba, and Casey (1985) and others (e.g., Gorman & Allison, 1996) and more recent approaches based on hierarchical linear models (e.g., Pustejovsky et al., 2014; Van den Noortgate & Onghena, 2003, 2008), make the assumption that model

errors are normally distributed and have constant variance. However, the most common types of outcome measurements used in SCD studies are frequency counts or percentages (Shadish & Sullivan, 2011), which may have characteristics that are not well modeled by normal distributions. Methodologists have therefore highlighted the need for statistical models that are more appropriate for the characteristics of outcome measures commonly used in SCDs (Shadish, 2014).

Recent work has begun to investigate other distributional models for SCD data. Most such work has focused on use of binomial distributions for outcomes in the form of percentages or proportions (e.g., Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Rindskopf, 2014) or poisson distributions for outcomes in the form of frequency counts (e.g., Declercq et al., 2018; Shadish et al., 2013a). Models based upon these distributions are well known and readily available in statistical software, yet these models also make strong assumptions of their own—assumptions that have not been thoroughly investigated in SCD data. A smaller amount of work has examined generalized linear and non-linear models with less strict distributional assumptions (Shadish et al., 2013a, 2013b; Swan & Pustejovsky, 2018), but these recent developments have yet to see many applications.

One of the most common approaches to measuring outcomes in SCDs is through systematic direct observation (SDO) of behavior (Ayres & Ledford, 2014; Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011), which entails observing a behavior in time and recording a quantitative summary of its features, using a system such as frequency counting, duration recording, momentary time sampling, or partial interval recording. SDO is considered a hallmark of single-case research (Horner et al., 2005) and a robust literature exists on the properties of different SDO procedures (Lane & Ledford, 2014; Pustejovsky & Runyon, 2014).

Despite the importance of SDO, only a few studies have considered the implications of SDO procedures for statistical modeling of SCDs. A recent exception is Yoder, Ledford,

Harbison, and Tapp (2018), who compared two different methods of estimating behavioral frequency from partial interval data. In closely related work, Pustejovsky and Swan (2015) developed effect size estimation methods specifically tailored for partial interval data. Finally, Pustejovsky (2018) used a model called the alternating renewal process (ARP) to simulate behavioral data and study the effects of different SDO procedures on effect size magnitude. The ARP model is distinctive because its formulation emulates the physical process of observing a behavior and recording SDO data. As a result, the model is very flexible and can be used to generate frequency counting data, duration recording, or interval recording data. However, such flexibly comes at the cost of substantial additional complexity. Further investigation is therefore warranted to determine whether the added complexity is necessary and appropriate for modeling real SCD data.

Monte Carlo simulations

Monte Carlo simulation is one of the primary tools used by methodologists to assess the performance of new statistical methods. Monte Carlo simulation involves simulating artificial data based on a model with known characteristics, applying one or more statistical procedures, and comparing the resulting estimates with the true features of the model (Morris, White, & Crowther, 2019). Repeating this process many times allows researchers to investigate or compare the performance characteristics (such as bias, accuracy, or Type I error) of different procedures.

Monte Carlo simulations have been used extensively in developing and evaluating methods for analysis of SCD data (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013; Pustejovsky et al., 2014). Simulations are particularly advantageous for studying the robustness of statistical methods because artificial data can be generated based on a model that does not conform to the assumptions of the procedures being studied (e.g., Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2016; Petit-Bois, Baek, Van den Noortgate, Beretvas, &

Ferron, 2015). Simulations have even been used to examine the performance of analytic methods that are not based on specific distributional assumptions, such as the non-overlap effect sizes for SCDs (e.g., Tarlow, 2017; Pustejovsky, 2018). However, Monte Carlo simulations are inherently limited by the models and conditions used to generate artificial data. In order for a simulation to yield findings that are relevant in practice, it must be based on a data-generating model that captures essential features of data that arise in real studies.

The present study

As the variety of statistical models for SCDs expands, and as researchers assess and compare methods using Monte Carlo simulations, it is critical that these efforts be informed by the features of real SCD data. In the present study, we address this need by surveying the characteristics of empirical SCD data, focusing in particular on the study designs, procedures used to collect outcome data, and distributional features of the resulting data.

We limit the scope of the survey in three important ways. First, we survey SCD studies included in extant systematic reviews of the single-case literature. This approach helps to ensure that included studies investigated interventions and outcomes for which it is plausible that researchers would be interested in applying statistical models (including effect sizes and meta-analysis, in particular). In contrast, previous reviews of the characteristics of SCDs have surveyed all available SCD studies published in certain journals (Hammond & Gast, 2010; Smith, 2012) or in a single year (Shadish & Sullivan, 2011) and therefore may have included idiosyncratic studies that would never be subjected to statistical analysis. Second, we focus on SCDs that measured behavioral outcomes using systematic direct observation, excluding SCDs studies focused solely on academic performance or other types of outcomes, where measurement considerations may be quite different. Third, we limit the analysis to SCD data from initial baseline phases only, so as to avoid the complexity of modeling changes in outcomes after interventions are

introduced. Including data from intervention phases would necessitate making further assumptions about how to best quantify changes in outcomes between baseline and intervention phases—important questions, but ones that warrant separate investigation. Under these limitations, the present study is guided by three research questions:

1. What are the basic features of SCDs included in extant systematic reviews of the single-case literature?
2. In SCDs included in extant systematic reviews, which procedures are used to measure behavioral outcomes?
3. For the most common types of measurements, what are the distributional features of baseline data in such SCDs?

The remainder of the paper proceeds as follows. In the next section, we briefly review several distributional models that have been proposed for SCD data and illustrate the key distinguishing features of the models. The following section details the sample of included studies, coding procedures, and data analysis methods. We then present results, followed by a discussion of limitations and implications.

Distributional models

Some of the most well-known statistical methods for SCD data are based on regression models with errors that are assumed to be normally distributed and homoskedastic (Center et al., 1985; Gorman & Allison, 1996). Methodologists have also developed heirarchical linear models for analyzing individual SCDs (Van den Noortgate & Onghena, 2003), estimating effect size indices (Pustejovsky et al., 2014), or synthesizing data across multiple SCDs (Van den Noortgate & Onghena, 2008). These models involve assumptions similar to single-level regression models, including normality and homoskedasticity of the lowest-level errors.

Because most SCD outcome data come in the form of frequency counts or proportions, Shadish (2014) argued for the importance of considering other, non-normal distributional models for SCD data, such as the poisson distribution for frequency counting data and the binomial distribution for proportions. The poisson distribution has long been used as a model for frequency counts of animal or human behavior observed over time (e.g., Altmann, 1974). For example, Shadish et al. (2013a) used the poisson distribution to model data from DiCarlo and Reid (2004), who measured the number of initiations of independent pretend-play behaviors of young children with disabilities. The binomial distribution is useful as a model for a variable representing the number of occurrences out of a fixed, known number of trials, where each trial is independent of the others. For example, a binomial distribution might be used to model the number of times a participant looks at an adult instructor in response to a bid for joint attention, as in Taylor and Hoch (2008). Researchers have begun to develop models for SCD data based on these distributions, such as the generalized linear mixed models described by Shadish et al. (2013a), Moeyaert et al. (2014), and Rindskopf (2014). Recently, Declercq et al. (2018) used Monte Carlo simulations to investigate the robustness properties of hierarchical linear models with normal error distributions when the true data-generating process involved poisson distributions.

In models such as these, there are at least two important considerations for choosing distributional assumptions. One consideration is the mean level of the outcome, which influences the extent to which poisson or binomial distributions are distinct from a normal distribution. In particular, a poisson distribution with a large mean is very closely approximated by a normal distribution (Johnson, Kemp, & Kotz, 2005). Similarly, a binomial distribution with mean near the middle of its range (i.e., mean near 50%) or with a large number of trials will also closely resemble a normal distribution (Johnson et al., 2005). Outside of these conditions, however, the poisson and binomial distributions become more distinct from the normal distribution. Consequently, models based on normal error

assumptions might be less robust when the outcomes are counts with low mean or percentage data based on a small number of trials.

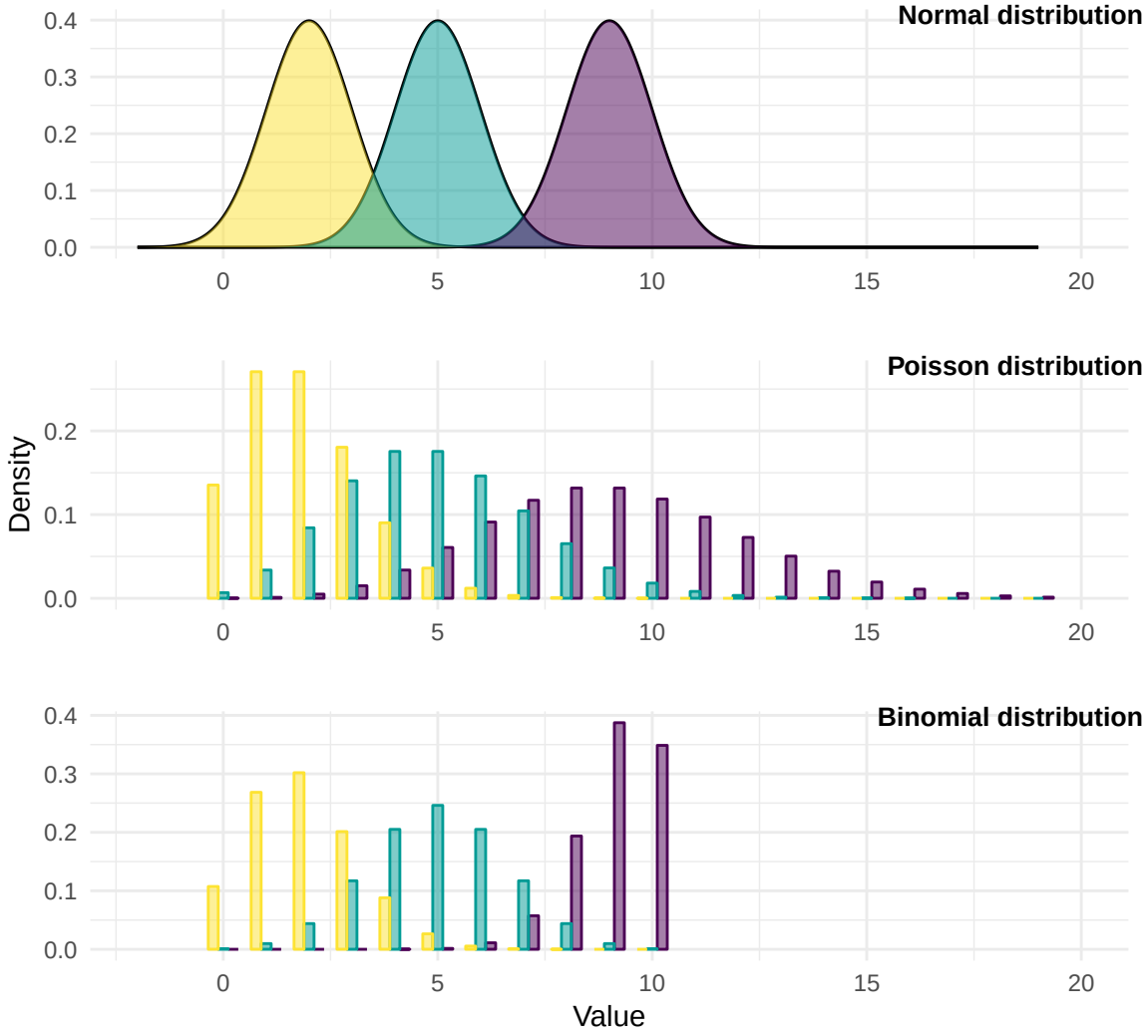

Figure 1. Distributions with means of 2, 5, and 9. The binomial distribution has size 10.

A second important consideration with these distributions is how the variability of the distribution is related to its mean. Figure 1 depicts each of these three distributions, each with mean values of 2, 5, or 9. Assuming homoskedasticity, the mean and variance of the normal distribution are unrelated, so that changes in the mean do not affect the degree of variability. In contrast, the poisson distribution has the property that its variance is equal to its mean. Consequently, a poisson distribution with a larger mean will necessarily also have higher variability. The middle panel of Figure 1 illustrates this relationship,

where it can be seen that the distribution with a mean of 9 has a wider spread than those with smaller mean values. The variability of a binomial distribution is also related to its mean, but in a different way than with the poisson. Letting μ denote the mean of a distribution, a binomial distribution with size K has the property that its variance is equal to $\mu(K - \mu)/K$. Thus, a binomial distribution has maximal variance when its mean is equal to $K/2$, with variance decreasing towards zero as the mean value approaches 0 or K.

In practice, it is quite common for empirical data to exhibit some mean-variance relationship, but not necessarily the exact relationships entailed by poisson or binomial distributions. Generalized linear models based on the quasi-likelihood framework (Fox, 2008; McCullagh & Nelder, 1989) provide one way to model count or proportion data under weaker distributional assumptions. Quasi-likelihood models are specified in terms of a mean-variance relationship that involves an additional parameter ϕ, known as the dispersion. For a count outcome P, the quasi-poisson model assumes that the variance is proportional—but not strictly equal—to the mean:

$$\text{Var}(Y) = \phi\mu,$$

where $\mu = \text{E}(Y)$. If $\phi = 1$, then the outcome follows a poisson distribution exactly; otherwise, the outcome is said to be *over-dispersed* (if $\phi > 1$) or *under-dispersed* (if $0 < \phi < 1$). Similarly, a quasi-binomial model assumes that the variance of the outcome is a quadratic function of its mean, which may be proportionally larger or smaller than the binomial variance:

$$\text{Var}(Y) = \phi\mu(K - \mu)/K.$$

Fox (2008) recommended using these quasi-likelihood models routinely due to the prevalence of over- and under-dispersion in empirical data. However, the extent and range of dispersion in SCD data remains to be explored.

One alternative to the conventional poisson or binomial distributions is the

alternating renewal process (ARP) model (Pustejovsky, 2015; Rogosa & Ghandour, 1991). The ARP model is formulated to describe a behavior that occurs episodically over time, such as bouts of stereotypy of self-injurious behavior during a therapy session (e.g., Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007). Under the model, it is assumed that the lengths of time in between bouts of behavior and the lengths behavioral episodes follow some probability distributions (such as exponential or gamma distributions). The stream of behavior is observed for a specified length of time, and an outcome is calculated by applying an SDO procedure such as frequency counting, momentary time sampling, or partial interval recording. Rogosa and Ghandour (1991) studied the ARP as a model for classroom behaviors such as teacher praise or time that a student is engaged or on-task.

A primary advantage of the ARP is its flexibility. Because it is a model for a stream of behavior, it can be used to simulate the effects of different SDO procedures, such as comparing momentary time sampling to partial interval recording or comparing data based on longer or shorter periods of observation (Meany-Daboul et al., 2007; Pustejovsky & Runyon, 2014). The ARP can also be used to simulate data that approximately conform to the assumptions of the quasi-poisson or quasi-binomial models, with varying degrees of dispersion. To illustrate, Figure 2 depicts three sets of frequency counting data simulated based on an alternating renewal process. Each distribution has a mean of 5, but the distributions have different spreads due to varying the degree of dispersion. The distribution with dispersion of 1 is identical to the poisson distribution with mean 5, as depicted in the middle panel of Figure 1.

One of the main challenges in using the ARP is its complexity. Although its formulation as a model for a stream of observed behavior emulates the physical process of SDO, this also makes it difficult to study analytically. The distributions of summary measurements (such as frequency counts or proportions of intervals) do not generally have simple mathematical forms. The model can nonetheless be used to simulate artificial data based on assumptions about behavior streams and measurement procedures (Pustejovsky

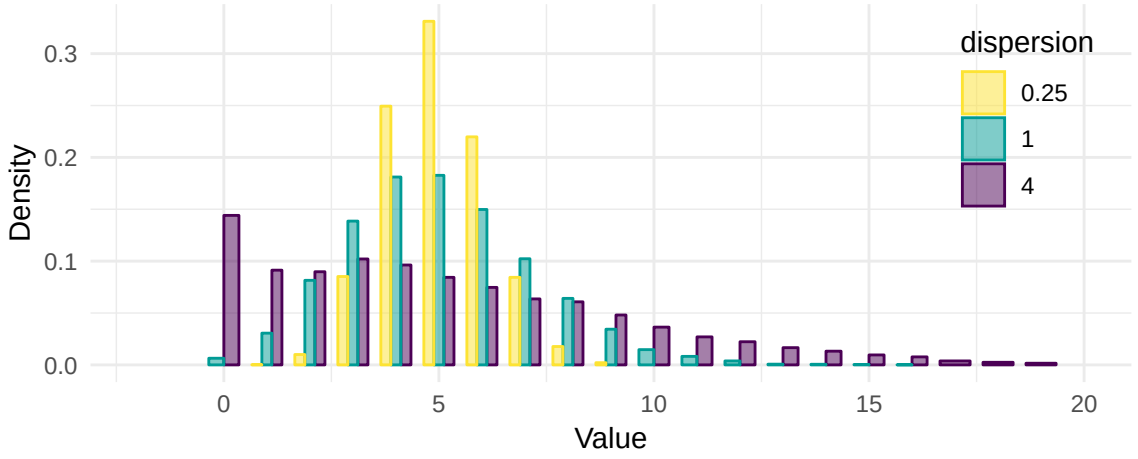

Figure 2. Frequency count data simulated from alternating renewal process with mean of 5 and varying levels of dispersion. Dispersion of 1 corresponds to a poisson distribution.

& Runyon, 2014). However, if Monte Carlo simulations based on the ARP are to be relevant for practice, the assumptions required of the simulation model should be informed by the features of actual SCD studies, including details of how outcomes are measured and the distributional properties of those outcomes. One of the goals of the present study is to assemble and summarize such details, so that future methodological studies can make more realistic, empirically informed assumptions.

Methods

Data sources

We drew on data from seven completed systematic reviews of SCD studies with behavioral outcomes. Table 1 describes characteristics of these reviews, including the populations, interventions, and outcomes of included primary studies, as well as whether primary studies in each review were screened based on the What Works Clearinghouse (WWC) standards for SCDs (Kratochwill et al., 2013). Shogren, Faggella-Luby, Bae, and Wehmeyer (2004) synthesized the literature on choice-making interventions for students with autism. Gage, Lewis, and Stichter (2012) is a systematic review examining the

Table 1
Characteristics of included systematic reviews

Review	Authors	Population	Interventions	Outcomes	Screening based on WWC?
Choice-making	Shogren, Faggella-Luby, Bae, & Wehmeyer (2004)	Individuals with disabilities	Choice-making opportunities	Disruptive behavior, on-task behavior, social skills	No
FBA	Gage, Lewis, & Stichter (2012)	Students with/at risk for emotional/ behavioral disorders	Functional behavioral assessment	Problem behavior	No
Group contingencies	Maggin, Pustejovsky, & Johnson (2017)	Students with challenging behavior	School-based group contingencies	Social interaction, academic engagement, disruptive behavior	Yes
Object play	Barton, Sweeney, & Gossett (2016)	Young children with disabilities	Least-to-most prompting, positive reinforcement	Object play	Yes
Peer management	Dart, Collins, Klingbeil, & McKinley (2014)	Students	Peer management interventions in school settings	Disruptive behavior, on-task behavior, social skills	No
PRT	Verschuur, Didden, Lang, Sigafoos, & Huskens (2014)	Individuals with ASD	Pivotal response training	Self-initiations, communication/language	No
Social skills	Ledford, King, Harbin, & Zimmerman (2016)	Individuals with ASD	Antecedent social skills	Pro-social behaviors	Yes

effectiveness of functional behavioral assessment (FBA) based interventions for students with or at risk for an emotional or behavioral disorder. Maggin, Pustejovsky, and Johnson (2017) examined school-based group-contingency interventions intended to reduce problem behaviors. Barton, Sweeney, E., and Gossett (2016) reported a systematic review of interventions intended to increase the frequency or enhance the quality of pro-social play behaviors. Dart, Collins, Klingbeil, and McKinley (2014) included studies on peer-management interventions targetting non-academic behaviors in school settings. Verschuur, Didden, Lang, Sigafoos, and Huskens (2014) synthesized studies on pivotal response treatment for children with autism spectrum disorders. Ledford, King, Harbin, and Zimmerman (2016) reviewed studies on social skills interventions for children with autism. In the three most recent systematic reviews (Barton et al., 2016; Ledford et al., 2016; Maggin et al., 2017), primary studies were screened and included in the review only if they met the WWC standards with or without reservations. In the remaining reviews, inclusion criteria were defined in terms of population, intervention, and outcome characteristics, but studies were not excluded based on methodological quality criteria.

Coding

The first authors of six of the seven included studies shared the underlying data from their reviews with us. In the case of Verschuur et al. (2014), we had access to a database from an incomplete systematic review that contained many of the same studies included in Verschuur et al. (2014). We then identified and coded the studies from the Verschuur et al. (2014) review that were not in the existing database. We compiled the data from all seven systematic reviews into a relational database and coded additional characteristics of all included studies.

Data from SCDs have a multi-level, hierarchical structure. Each primary study in the included reviews involved one or more participants. For some types of SCDs, such as multiple baselines across participants, there is a one-to-one correspondence between

participants and data series (i.e., the tiers of the design). However, it is also possible that multiple data series are measured on a given individual, such as in a multiple baseline across settings or in a treatment reversal design in which multiple outcomes are recorded for each session. In such cases, series are nested within partipant. Finally, each data series consists of data points collected across multiple sessions, so that sessions are nested within data series. We built a relational database that followed this hierarchical structure and coded characteristics at each of four levels: the study level, the participant level, the series level, and the session (data-point) level.

Study characteristics. We coded the type of design implemented in each study, following the typology described by Gast and Ledford (2018). Basic design types included AB (i.e., pre-experimental), multiple baseline across participants, multiple baseline across settings or behaviors, multiple probe across participants, multiple probe across settings or behaviors, treatment reversal, alternating treatment, adapted alternating treatment, or changing criterion. For designs that incorporated features of multiple design types, we classified the study according to the predominant basic type while also noting the adaptations or variations. For example, we classified a multiple baseline across participants where one tier of the design included treatment reversals as a multiple baseline across participants.

Participant characteristics. At the participant level, we extracted the pseudonym used to describe each case in the primary study. Additionally, some designs involved outcomes assessed on whole groups (such as an entire classroom of students), while other designs involved outcomes assessed at the individual level. We therefore classified the participants as individuals or aggregates.

Series characteristics. The majority of the coding pertained to series-level characteristics. For each data series, we coded several aspects of the outcome measurements procedures, including the class of measurement procedure, details of the measurement procedure, length of observation session, and outcome metric. Measurement

procedures fell under two broad categories: those for measuring free-operant behavior and those for measuring restricted-operant behavior. SDO procedures for free-operant behavior included continuous recording (i.e., durational recording), momentary time sampling (MTS), event counting, partial interval recording (PIR) and whole interval recording (WIR). Procedures for measuring restricted operant behavior included assessing successes on a fixed or variable number of stimuli/tasks, response latencies, and task check-lists. A few studies used outcomes measured using self-reported or clinician-reported rating scale measures, typically assessing psychological constructs. For interval recording methods, we coded the length of the observation intervals and the amount of recording time in between observation intervals (e.g., 15 s PIR, with 5 s space in between intervals).

For data series in which outcomes were measured through SDO of free-operant behavior, the length of the observation session may influence the variability of the measurements, with longer sessions expected to produce more stable measures. We therefore coded the reported length of observation sessions for data series measured using interval recording, momentary time sampling, event counting, or continuous recording. If sessions varied in length, we used the average length or mid-point of the range of session lengths as reported by study authors. Finally, we coded the metric in which the outcome was reported, so that we could standardized the format of outcomes for purposes of analysis. Possible metrics included percentages (on a scale of 0-100%), proportions (on a scale of 0-1), number of successes out of a fixed total, event counts with no natural upper limit, standardized rates of event frequency, or miscellaneous other metrics.

Session characteristics. In order to examine the distributional characteristics of baseline outcomes, we needed data on session-level outcome measurements, session sequence, and phases. For six of the seven reviews, the original authors provided session-level outcome data and phase information, which we cleaned, verified against graphs of the data in the source articles, and read into our database. For the Verschuur et al. (2014) review, we followed this process for a portion of the studies for which we had access

to outcome data. For the remaining studies, we extracted outcome data from the primary source articles using the Web Plot Digitizer tool (Rohatgi, 2015). We coded each individual data point extracted from a plot with a session number and a label for the phase of the study. The data provided from Dart et al. (2014) included only initial baseline phases.

Data analysis

Our main approach to data analysis involved calculating descriptive summary statistics, including percentages, means, medians, and inter-quartile ranges (IQR) for characteristics of the included studies, participants, and data series.

To analyze the features of the distribution of baseline outcomes, we restricted the sample in three ways. First, we limited the sample to data from initial baseline phases only. Second, we also limited the sample to series in which the initial phase of the study was a baseline (non-intervention). This resulted in the exclusion of 53 data series from 13 studies. Finally, we restricted our analysis to the two most common types of measurements used in single-case research: event frequency counts and proportion-based measures of free-operant behavior (i.e., PIR, WIR, MTS, and continuous recording).

For event frequency counts, we first converted all of the data series to a common metric. Specifically, for outcomes reported as standardized rates per minute, we multipled the outcome data points by the session length to obtain frequency counts. We then calculated the sample mean and sample variance of each baseline data series, examined the empirical distribution of baseline means, and created scatterplots of the sample mean and sample variance in order to investigate their relationship.

For measures of free-operant behavior reported as percentages or proportions, we first converted all of the series to the common metric of proportions. We then calculated the sample mean and sample variance of each baseline data series. Because different series were measured using different numbers of intervals and the variance of a proportion depends on

the number of intervals, sample variances had to be re-scaled. We put the them on a common metric by multiplying each sample variance by the corresponding number of intervals used to measure the outcome for that data series. For scaling purposes, we treated continuous recording data as equivalent to 5 s momentary time sampling, so that the number of intervals was equal to 12 times the length of the observation session in minutes. If outcomes exactly followed a binomial distribution, then the expected scaled variance would be equal to $\mu(1-\mu)$, which has a maximum value of 0.25 when $\mu = .5$. Thus, scaled sample variances greater than $\mu(1-\mu)$ correspond to over-dispersion relative to a binomial distribution. Just as with the event counting data, we examined the distribution of baseline mean values and created scatterplots of the sample mean and sample variance for data series measured using each procedure.

We carried out all calculations and created figures and tables using R (Version 3.6.0; R Core Team, 2019). R packages used in creating the figures included ggplot2 (Wickham, 2016), ggridges (Wilke, 2018), cowplot (Wilke, 2019), and colorspace (Zeileis, Hornik, & Murrell, 2009). The complete raw data and R code for replicating our analysis are available at https://osf.io/n9jud/.

Results

Study design characteristics

Table 2 reports the distribution of study design types both overall and by review. Across reviews, nearly half (48%) of the 303 primary studies used multiple baselines across participants. The next most common design types were treatment reversals (34%) and multiple baselines across behaviors (10%), with the remaining types used in less than 10% of studies. It is notable that the frequency of different design types varied by systematic review. In the pivotal response training and social skills reviews, across-participant multiple baseline and multiple probe designs were predominant, and the object play review

Table 2
Distribution of study designs by review

Design	Overall (n=303)	Choice-making (n=13)	FBA (n=67)	Group contingencies (n=40)	Object play (n=11)	Peer management (n=29)	PRT (n=31)	Social skills (n=112)
MBP	145 (48%)	2 (15%)	15 (22%)	11 (28%)	8 (73%)	15 (52%)	27 (87%)	67 (60%)
TR	102 (34%)	10 (77%)	39 (58%)	28 (70%)	-	11 (38%)	3 (10%)	11 (10%)
MBB	31 (10%)	1 (8%)	9 (13%)	1 (2%)	-	3 (10%)	-	17 (15%)
MPP	13 (4%)	-	-	-	3 (27%)	-	-	10 (9%)
Other	12 (4%)	-	4 (6%)	-	-	-	1 (3%)	7 (6%)

Note: MBP = Multiple baseline across participants. TR = Treatment reversal. MBB = Multiple baseline across behaviors or settings. MPP = Multiple probe across participants. Other designs include AB, alternating treatment, adapted alternating treatment, and multiple probes across behaviors or settings. FBA = Functional behavior assessment. PRT = Pivotal response training.

consisted entirely of such designs. In contrast, treatment reversal designs comprised a large majority of primary studies in the choice-making, functional behavior assessment, and group contingency reviews.

Table 3 reports further basic features of the included studies. Across reviews, the majority of studies included between 2 and 4 unique participants (mean = 3.1, median = 3). The distribution of the number of participants was generally consistent across the reviews, although there were a few studies that included a larger number of participants. Across reviews, 8% of studies included six or more participants, including a doctoral dissertation by Thorne (2005) with 12 participants and a study by Laski, Charlop, and Schreibman (1988) with 9 parent-child dyads.

On average across reviews, one or two data series were measured per participant, again with little variation across most reviews. The exception was the pivotal response training review, which included an average of 3.10 data series per participant. This review also included the largest number of participants per study (mean = 4.4, median = 3).

Across reviews, initial baseline phases included an average of data points per series (median = 7, IQR = 5-12). As depicted in Figure 3, the distribution of baseline phase lengths was strongly right-skewed, with the longest baseline phase consisting of sessions. Of 1765 data series that began with a baseline phase, 112 series (from 24 unique studies) had 30 or more observations in the initial phase.

Table 3
Study characteristics by review

		Participants per study			Series per participant			Baseline sessions per series		
Review	Studies	Mean	Median	IQR	Mean	Median	IQR	Mean	Median	IQR
Overall	303	3.1	3	2-4	2.0	1	1-2	11.1	7.0	5-12
Choice-making	13	2.5	3	1-3	2.3	2	1-3	7.4	5.5	4-9
FBA	67	2.1	2	1-3	1.5	1	1-2	8.6	7.0	4-10
Group contingencies	40	3.3	3	2-4	1.5	1	1-2	7.4	5.0	4-9
Object play	11	3.8	3	3-4	1.3	1	1-2	11.3	9.5	5-12
Peer management	29	3.0	3	2-3	1.6	1	1-2	10.8	8.0	5-13
PRT	31	4.4	3	3-5	3.1	2	1-5	9.5	6.0	4-11
Social skills	112	3.2	3	3-4	2.0	2	1-2	13.9	9.0	6-16

Note: IQR = Inter-quartile range. FBA = Functional behavior assessment. PRT = Pivotal response training.

Measurement procedures

Table 4 reports the distribution of procedures used to measure outcomes, both overall and in each of the seven reviews. Across reviews, the most common measurement procedure was PIR (36% of data series), followed by event counting (23%) and successes out of a fixed number of trials (20%). Notably, PIR was the most common procedure in five out of the seven reviews and was used in more than half of all data series in the functional behavior assessment and pivotal response training reviews. Among the 650 data series measured using PIR, the majority used 10 s interval lengths (61%), followed by 30 s intervals (18%). Only 6% of data series used PIR with interval lengths of less than 10 s. WIR and MTS interval lengths were similar to those used with PIR. Specifically, 10 s intervals were used in 84% of the 69 data series measured using WIR and 47% of the 86 data series measured using MTS.

Among procedures for measuring restricted operant behavior, successes out of a fixed or variable number of trials were common overall, but their use was not uniformly frequent across all reviews. Rather, these procedures were most common in the reviews of social skills training, where they comprised 45% of all data series, and pivotal response training, where they comprised 24% of all data series. In the other reviews, these procedures were rarely used.

Table 4
Distribution of measurement procedures by review

Procedure	Overall	Choice-making	FBA	Group contingencies	Object play	Peer management	PRT	Social skills
partial interval recording	650 (36%)	22 (31%)	135 (67%)	52 (26%)	26 (48%)	64 (44%)	247 (58%)	104 (14%)
event counting	419 (23%)	7 (10%)	13 (6%)	61 (31%)	21 (39%)	42 (29%)	61 (14%)	214 (30%)
success-fixed	363 (20%)	-	-	3 (2%)	3 (6%)	8 (6%)	72 (17%)	277 (38%)
success-variable	97 (5%)	11 (15%)	2 (1%)	-	-	4 (3%)	28 (7%)	52 (7%)
momentary time sampling	86 (5%)	1 (1%)	16 (8%)	66 (34%)	-	3 (2%)	-	-
continuous recording	80 (4%)	6 (8%)	16 (8%)	15 (8%)	-	3 (2%)	-	40 (6%)
whole interval recording	69 (4%)	10 (14%)	19 (9%)	-	-	15 (10%)	3 (1%)	22 (3%)
response latency	19 (1%)	-	-	-	-	-	6 (1%)	13 (2%)
task check-list	18 (1%)	10 (14%)	-	-	-	5 (3%)	-	3 (0%)
rating scale	17 (1%)	4 (6%)	1 (0%)	-	4 (7%)	-	8 (2%)	-

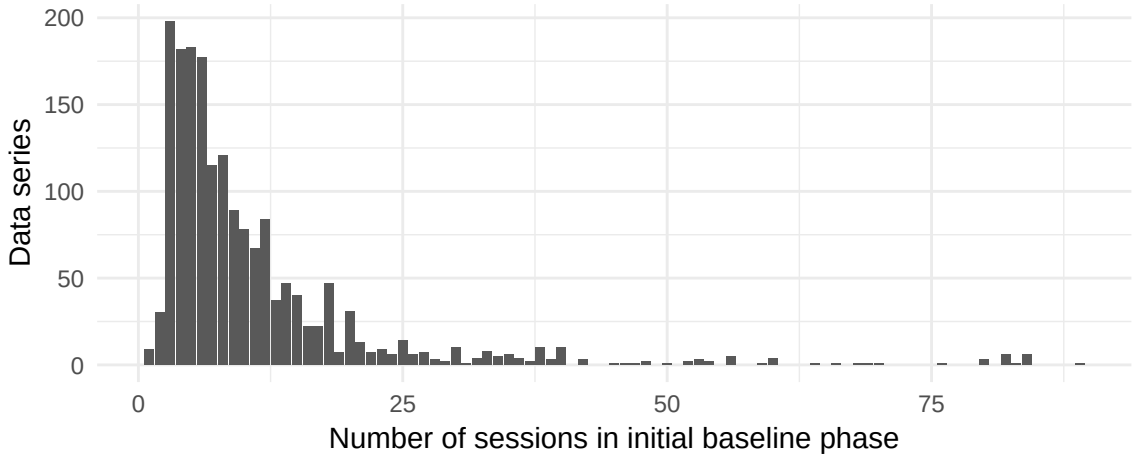

Figure 3. Distribution of initial baseline phase lengths across seven systematic reviews of SCDs.

When using SDO to measure free-operant behavior, session length may have an influence on the variability (or stability) of the measurements. Figure 4 depicts the distribution of session lengths among the 1233 data series that used SDO to measure free-operant behavior and where session length was reported. Across all reviews, sessions lasting longer than 30 minutes were uncommon, occurring in less than 5% of data series. Overall, observation sessions had a median length of 10 minutes (IQR = 10-15). In four out of the seven reviews, 10 minutes was also the modal session length. In the review of choice-making interventions, the median session length was somewhat longer at 15 minutes (IQR = 14-15), and still longer, at 20 minutes (IQR = 15-30) in the group contingency review.

Baseline outcomes: event frequency counts

Turning to the properties of baseline outcome measurements, we first examined the properties of event frequency counting data. In doing so, we distinguish between positive-valence outcomes (where increasing the behavior is desirable) and negative-valence outcomes (where decreasing the behavior is desirable), due to substantial differences in their mean frequency.

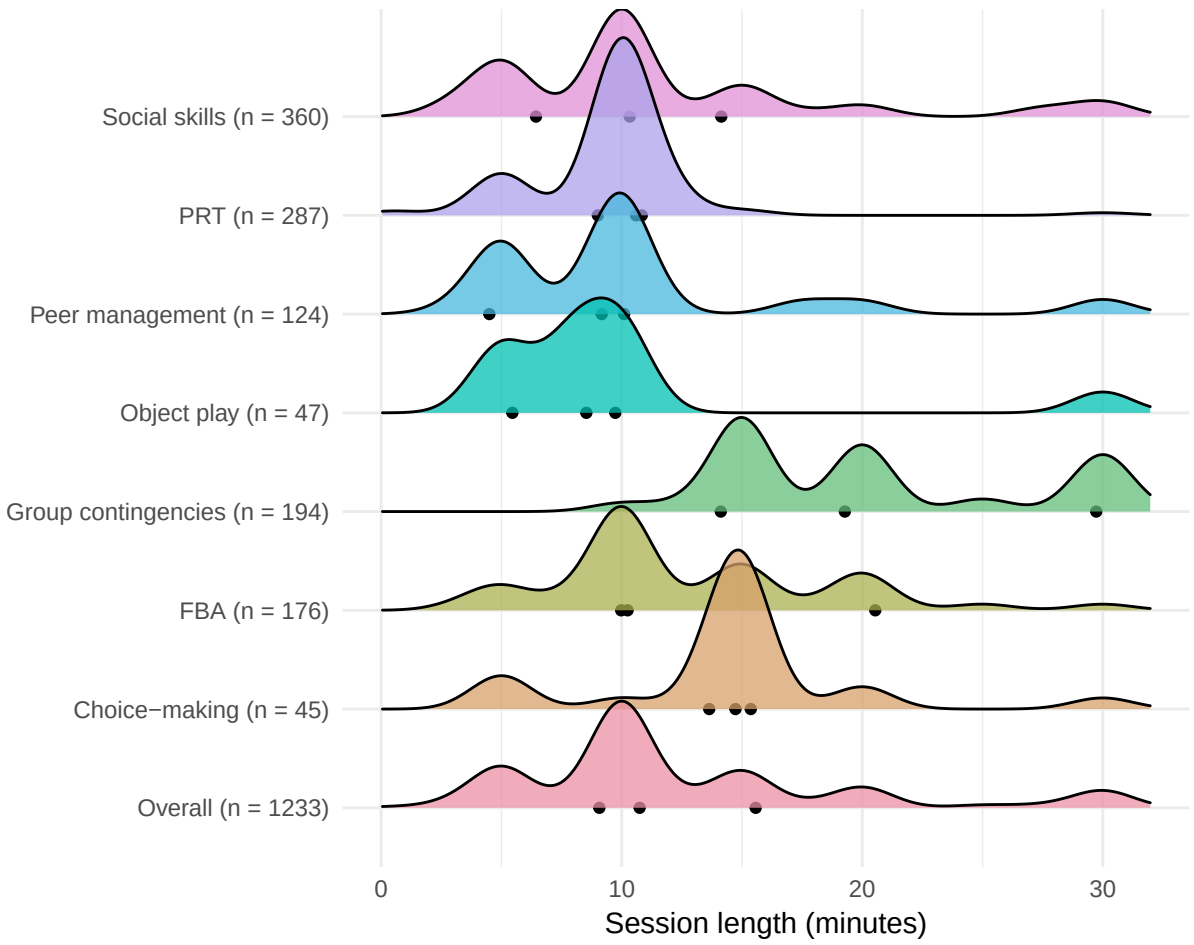

Figure 4. Distribution of session lengths (in minutes) in data series measuring free-operant behaviors. Curves represent smoothed kernel density estimates. Dots indicate the quartiles of the distribution for each review. Session lengths over 30 minutes are not depicted.

Figure 5 depicts the distribution of mean frequencies and scatterplots of the mean-variance relationship for positive-valence (left-hand plots) and negative-valence (right-hand plots) outcomes. Several trends are apparent. First, positive-valence outcomes have quite low average baseline frequency, with a median of 0.83 events per session (IQR = 0-3.4). Of the 311 data series with positive valence event count outcomes, fully 27% had mean baseline levels of 0.01 or less—effectively zero. Most data series with zero baselines came from the review of social skills interventions.

Second, compared to positive-valence outcomes, negative-valence outcomes have higher–and more variable–average baseline frequency. The median baseline frequency of

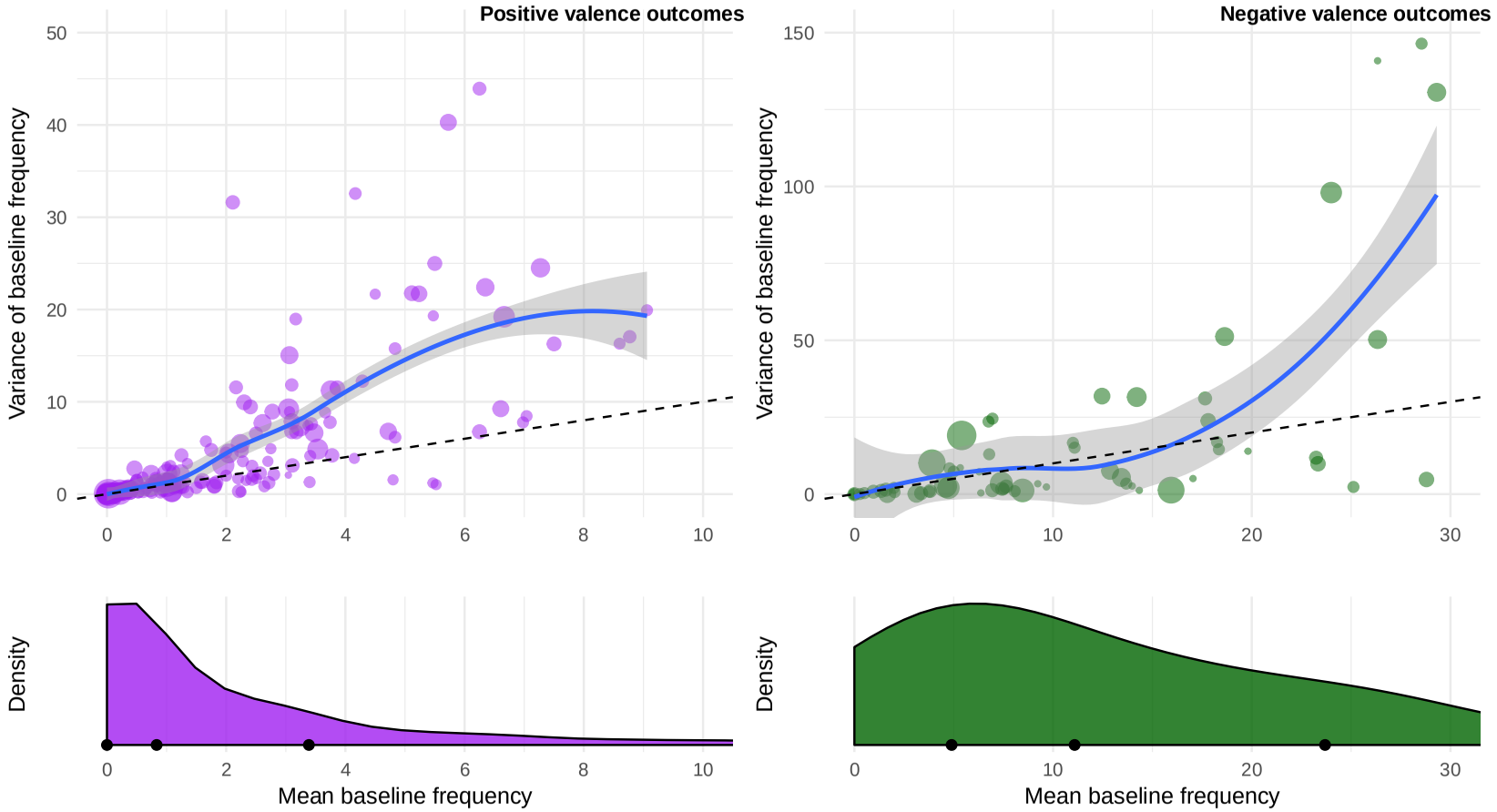

Figure 5. Scatterplots of variance versus sample mean baseline frequency for event count outcomes, with marginal distributions of mean baseline frequency. Left-hand plots (purple) depict positive-valence outcomes. Right-hand plots (green) depict negative-valence outcomes. Each point represents one data series, with size corresponding to baseline phase length. Dashed lines represent unit slopes, where variance is equal to mean. Blue curves depict local linear regressions of variance as a function of mean. Dots along the horizontal axis indicate the quartiles of the distribution.

these series was 11 events per session (IQR = 4.9- 24), and very few series had baseline levels near zero.

Third, both positive- and negative-valence frequency counts exhibited positive mean-variance relationships. Series with larger sample means tended to also have larger sample variances. This pattern suggests that distributions in which the mean and variance are connected (such as the poisson or ARP) may be useful in modeling or simulating single-case count data.

Fourth, for positive-valence outcomes, sample variances of the baseline series tended to exceed the sample means, indicating overdispersion of the distributions. For negative-valence outcomes, dispersion was around one for baseline frequencies of less than 15, but tended to exceed one for higher baseline frequencies. If outcomes followed a poisson distribution, for which the mean and variance are equal, we would expect that sample variances would be similar to sample means. However, across both types of outcomes, the patterns suggest that it may be useful to use distributions that have varying degrees of dispersion—and particularly, in which over-dispersion is possible—for modeling or simulating single-case data.

Baseline outcomes: proportions

In examining the distributions of outcomes measured as proportions, we focused on those that were measures of free-operant behavior, collected using PIR, MTS, continuous recording, or WIR. Figure 6 depicts the distribution of baseline mean outcomes for data series measured using each of these procedures, with the distribution of positive-valence outcomes on the left and negative-valence outcomes on the right. For both positive- and negative-valence outcomes, the distribution of mean baseline levels spanned the full range from 0 to 1. Beyond that, however, the two types of outcomes had distinctive distributions. Positive-valence outcomes have fairly low mean levels, with median levels ranging from 0.05

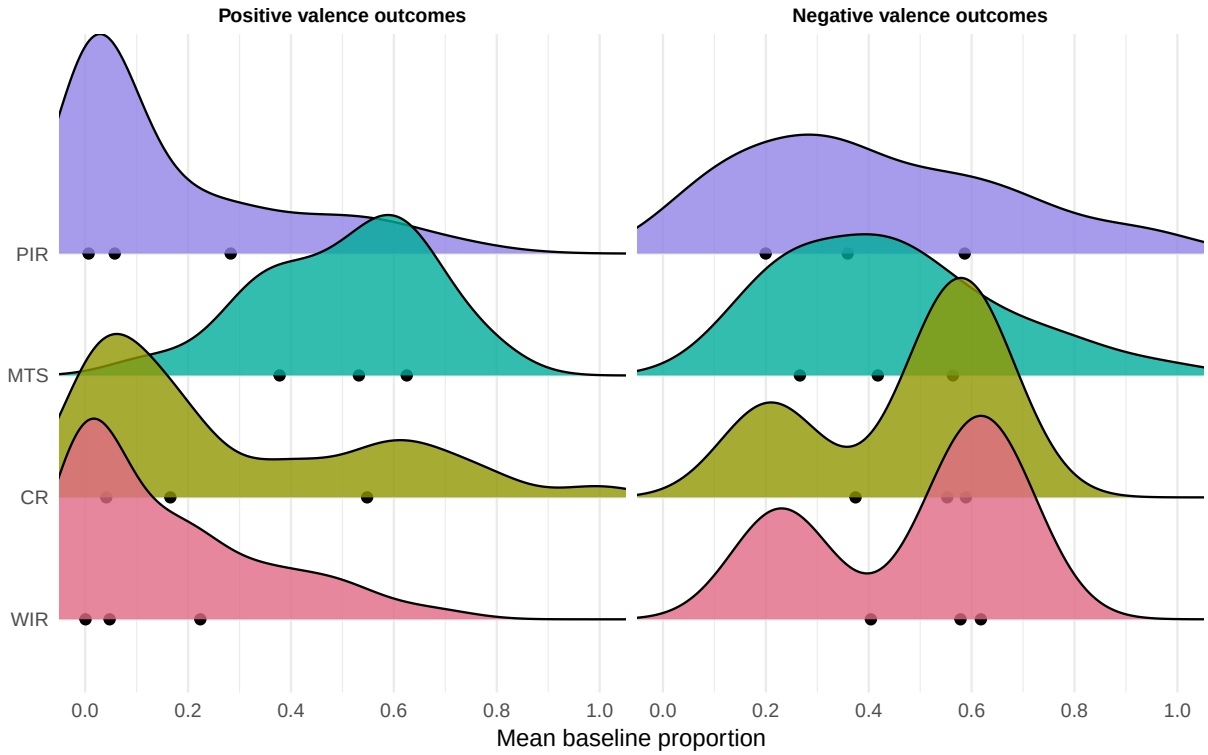

Figure 6. Distributions of mean baseline proportions for partial interval recording (PIR), momentary time sampling (MTS), continuous recording (CR), and whole interval recording (WIR) data series. Left-hand plots depict positive-valence outcomes. Right-hand plots depict negative-valence outcomes. Dots along the horizontal axis indicate the quartiles of the distribution.

for WIR and 0.06 for PIR to 0.17 for continuous recording. Surprisingly, the distribution of mean baseline outcomes measured using MTS was near the center of the range, with a median of 0.53 (IQR = 0.38-0.63). Across the 591 data series with positive-valence outcomes measured using any of the four procedures, 25% had mean baselines very close to zero (0.01 or less). In contrast, the bulk of the data series with negative-valence outcomes fell close to the center of the range. For instance, with the most commonly used procedure of PIR, the median baseline level was 0.36 (IQR = 0.20-0.59). Of the data series with negative-valence outcomes, none had a baseline level lower than 0.01.

Figure 7 includes scatterplots of the scaled sample variance versus the sample mean proportions for series measured using each of the four procedures. Positive- and

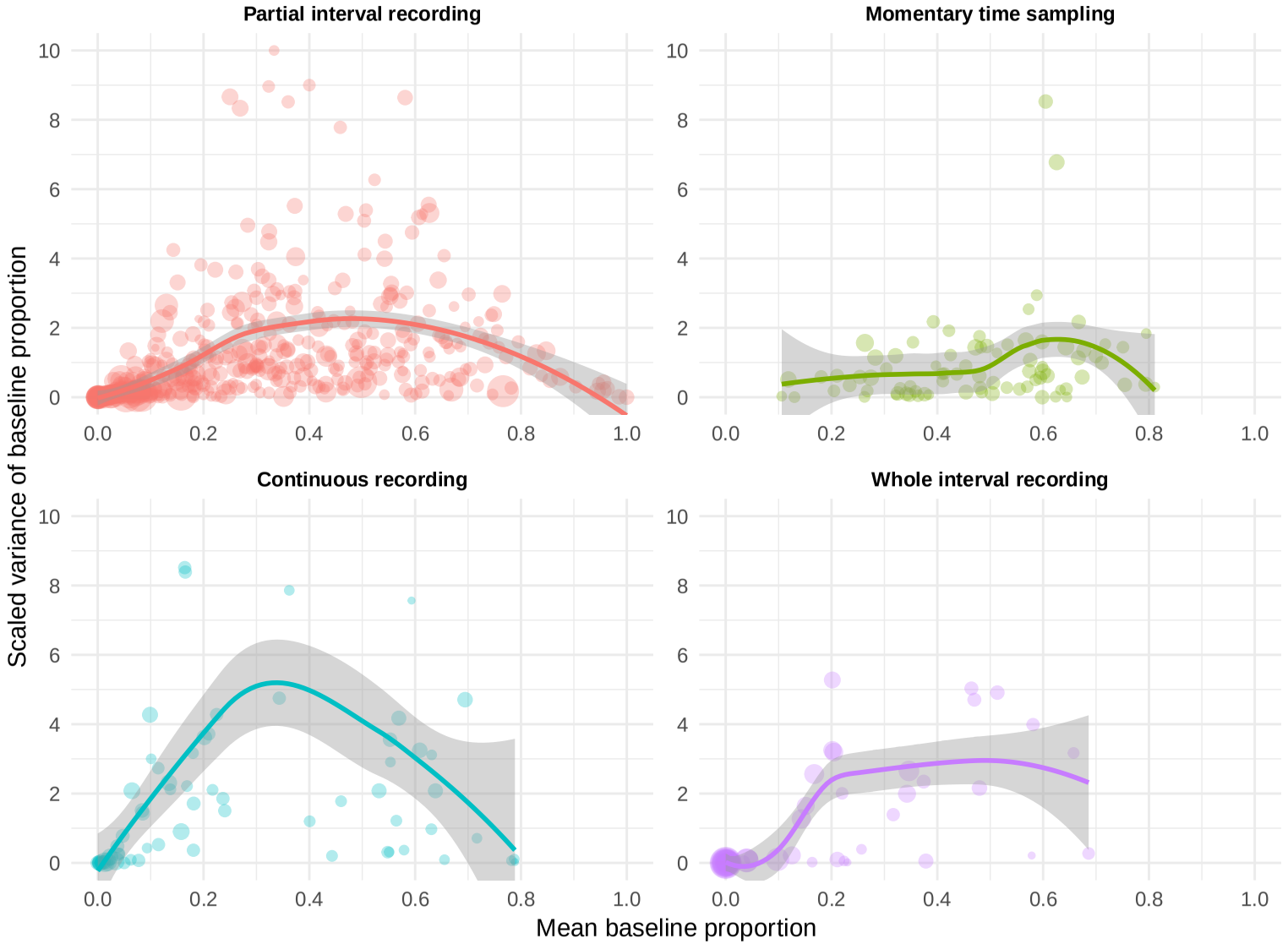

Figure 7. Scatterplots of sample variance versus sample mean baseline proportions for partial interval recording, momentary time sampling, continuous recording, and whole interval recording data series. Each point represents one data series, with size corresponding to baseline phase length. Curves depict local linear regressions of variance as a function of mean.

negative-valence outcomes are depicted in the same plot because they showed very similar patterns of results and because there were only a few series with negative-valence outcomes measured by WIR or CR. Just as with the frequency counting data, there are apparent functional relationships between the mean and variance of each type of outcome: larger scaled variances tend to occur at intermediate mean levels while smaller scaled variances occur at mean levels closer to the extremes of zero and one. If outcomes followed a binomial distribution, we would expect that sample variances would fall around $\mu(1-\mu)$, which reaches a maximum of 0.25 when $\mu = 0.5$. However, for each of the four measurement procedures, scaled sample variances tended to systematically exceed this level. For instance, PIR data exhibited a median dispersion of 3.9 times what would be expected for a binomial distribution (IQR = 1.4-8.0). This suggests that distributional assumptions that allow for overdispersion may be useful—and more realistic than the binomial distribution, which has dispersion of one—when modeling or simulating proportion data in SCDs.

Discussion

This paper has examined several aspects of the study designs, operational procedures, and outcome data in single-case designs, focusing on designs that are included in existing systematic reviews for behavioral interventions. Notably, the corpus of studies included in this survey is drawn from several fields, ranging from school psychology (Dart et al., 2014) to behavioral disorders (Gage et al., 2012) to autism and developmental disorders (Ledford et al., 2016; Verschuur et al., 2014). Although the seven included systematic reviews are only a convenience sample, we believe that the range of topics and settings of the included studies provides some reason to expect that our findings hold more broadly—generalizing to the literature base of SCD studies with behavioral outcomes that would be considered eligible for inclusion in systematic reviews. We would not, however, generalize to the single-case literature base on academic outcomes (e.g., Burke, Boon, Hatton, & Bowman-Perrott, 2015), teacher process outcomes, or other outcomes where systematic

direct observation of behavior is less commonly used.

Findings regarding the types and basic features of SCDs are largely consistent with past reviews of the single-case literature. Past systematic reviews of SCDs have also found that multiple baselines are the single most common type of SCD (Hammond & Gast, 2010; Shadish & Sullivan, 2011; Smith, 2012) and that baseline phases include an average of approximately 11 observations (Smith, 2012). These previous reviews of SCD characteristics all examined samples defined by inclusion in specific journals (Hammond & Gast, 2010; Smith, 2012) or publication in a single year (Shadish & Sullivan, 2011). Using a sample defined by inclusion in a systematic review, we have highlighted that the prevalence of different design types varies across systematic reviews, with treatment reversal designs playing a more prominent role in some areas (e.g., Maggin et al., 2017) than in others (e.g., Verschuur et al., 2014).

Consistent with past reviews on use of SDO measurement systems (Adamson & Wachsmuth, 2014), we found that use of partial interval recording is quite common—even used in two thirds of data series in the review of functional behavioral assessment (Gage et al., 2012). This pattern is troubling because it is well known that PIR does not provide a clearly interpretable measure of behavioral frequency or duration (Lane & Ledford, 2014). There are theoretical illustrations that use of PIR can create misleading patterns of results, even making a beneficial intervention appear to be harmful or vice versa (Pustejovsky & Swan, 2015), as well as some empirical evidence that use of PIR leads to exaggerated evidence of intervention efficacy (Radley, O'Handley, & Labrot, 2015). Some recent work has begun to develop statistical methods tailored specifically to PIR data (Pustejovsky & Swan, 2015; Yoder et al., 2018), but further research efforts in this direction are needed.

Previous work has noted that outcomes in SCD studies are commonly in the form of counts, percentages, or proportions (Shadish & Sullivan, 2011) and suggested use of poisson and binomial distributions in models for such data (Shadish, 2014). In light of the range of

average baseline levels—as well as the apparent mean-variance relationships—observed in the present review, further development of models for non-normal outcomes does seem warranted. However, poisson and binomial distributions may be insufficiently flexible to capture the features of real SCD data. We found that baseline frequency count data and proportion data exhibited a wide range of dispersions, often exceeding the levels of variance that would be expected under poisson or binomial models.

One alternative to such distributional models is the alternating renewal process, a highlight flexible model that emulates the physical process of SDO and that can be used to simulate data with over- or under-dispersion. Use of the ARP does, though, require specifying the length of observation sessions and the type of procedure used to record data. Such assumptions can be informed by the empirical distribution of these procedural choices, as we have reported. Other alternative models that could be investigated further include the beta-binomial, double binomial, and double poisson distributions (Efron, 1986; Johnson et al., 2005), all of which can accomodate varying degrees of dispersion. Future simulation studies exploring such models can use the distribution of mean levels and dispersions reported here to inform the range of parameters examined.

Findings from this review do have several important limitations. As we have noted, our analysis is based on a convenience sample of systematic reviews of single-case studies with behavioral outcomes, and does not necessarily generalize to studies in other research areas or with other types of outcomes. Furthermore, our analytic approach was limited to descriptive statistics, without quantifying the extent of uncertainty in the findings. This issue is most salient for the results regarding outcome distributions, where part of the observed variation in the distribution of outcome levels and dispersions is due purely to sampling variation. In ongoing work, we are exploring generative statistical models, such as generalized additive models for location and scale (Rigby & Stasinopoulos, 2005) that would make it possible to parse out study-level, participant-level, and session-level (sampling) variation.

Another limition of our analytic approach is that we have focused only on the two simplest features—the mean and variance—of the baseline outcome distribution. Other aspects of baseline data patterns—including time trends, auto-correlation, and inter-rater agreement—are also important to consider when developing models for SCDs (Shadish, 2014). Past surveys of the degree of serial dependence in single-case data (e.g., Shadish & Sullivan, 2011) have found moderate levels of positive auto-correlation, but the approach in these analyses is based on models involving normal error distributions. Given the prevalence of non-normal count and proportion outcomes, combined with the fairly short data series of most single-case data, estimating and modeling auto-correlation is challenging and further methodological research is needed. Further, assessing inter-rater agreement and attaining high levels are considered important markers of rigorous single-case research (Horner et al., 2005; Kratochwill et al., 2013), but little work has explored how to account for inter-rater agreement in statistical models for single-case data.

In examining the outcome measurement procedures most commonly used in the included SCDs, one particularly notable pattern was that the systematic reviews varied in the types of measurement procedures most commonly used. To the extent that the choice of effect size measures and statistical models for SCDs depends on the properties of the outcome measurements, different methods might well be needed for reviews in different areas. Rather than aiming to identify one single effect size measure or statistical model that is ideally suited for analysis or meta-analysis of SCDs, a better goal may be to develop methods that are tailored to the properties of the data under analysis. In doing so, researchers should carefully attend to the properties of the study designs and measurement procedures that were used to collect the data.

References

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

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*(3/4), 227–267.

Ayres, K., & Ledford, J. R. (2014). Dependent measures and measurement systems. In D. L. Gast & J. R. Ledford (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (pp. 124–153). New York, NY: Routledge.

Barton, E. E., Sweeney, E., & Gossett, S. (2016). A review of object play interventions for young children with disabilities. *Manuscript in Preparation.*

Burke, M. D., Boon, R. T., Hatton, H., & Bowman-Perrott, L. (2015). Reading Interventions for Middle and Secondary Students With Emotional and Behavioral Disorders: A Quantitative Review of Single-Case Studies. *Behavior Modification*, *39*(1), 43–68. doi:10.1177/0145445514547958

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-Subject Experimental Design for Evidence-Based Practice. *American Journal of Speech-Language Pathology*, *21*(4), 397. doi:10.1044/1058-0360(2012/11-0036)

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.

Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review*, *43*(4), 367–384.

Declercq, L., Jamshidi, L., FernÃndez-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron,

J. M., & Van den Noortgate, W. (2018). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods*. doi:10.3758/s13428-018-1091-y

DiCarlo, C. F., & Reid, D. H. (2004). Increasing pretend toy play of toddlers with disabilities in an inclusive setting. *Journal of Applied Behavior Analysis*, *37*(2), 197–207. doi:10.1901/jaba.2004.37-197

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, *81*(395), 709–721.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*(2), 372–84. doi:10.3758/BRM.41.2.372

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Los Angeles: Sage.

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*, 55–77.

Gast, D. L., & Ledford, J. R. (2018). *Single case research methodology: Applications in special education and behavioral sciences*. New York, NY: Routledge.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and Analysis of Single-Case Research* (pp. 159–214). Mahwah, NJ: Lawrence Erlbaum.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental*

Disabilities, *45*(2), 187–202.

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., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91–125). Washington, DC, US: American Psychological Association. doi:10.1037/14376-004

Jamshidi, L., Heyvaert, M., Declercq, L., FernÃndez-Castilla, B., Ferron, J. M., Moeyaert, M., ... Van den Noortgate, W. (2018). Methodological quality of meta-analyses of single-case experimental studies. *Research in Developmental Disabilities*, *79*, 97–115. doi:10.1016/j.ridd.2017.12.016

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (Vol. 444). John Wiley & Sons.

Kahng, S., Ingvarsson, E. T., Quigg, A. M., Seckinger, K. E., & Teichman, H. M. (2011). Defining and measuring behavior. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 113–131). New York, NY: Guilford Press.

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38.

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

Lane, J. D., & Ledford, J. R. (2014). Using interval-based systems to measure behavior in early childhood special education and early intervention. *Topics in Early Childhood Special Education*. doi:10.1177/0271121414524063

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, *21*(4), 391–400.

Ledford, J. R., King, S., Harbin, E. R., & Zimmerman, K. N. (2016). Antecedent social skills interventions for individuals with asd: What works, for whom, and under what conditions? *Focus on Autism and Other Developmental Disabilities*, 1088357616634024.

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., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education*, 0741932517716900. doi:10.1177/0741932517716900

Manolov, R., & Moeyaert, M. (2017). How Can Single-Case Data Be Analyzed? Software Resources, Tutorial, and Reflections on Analysis. *Behavior Modification*, *41*(2), 179–228. doi:10.1177/0145445516664307

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London, UK: Chapman & Hall.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, *40*(3), 501–514. doi:10.1901/jaba.2007.40-501

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. doi:10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A monte carlo simulation study. *Multivariate Behavioral Research*, *48*(5), 719–748. doi:10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A monte carlo simulation study. *The Journal of Experimental Education*, *84*(3), 473–509. doi:10.1080/00220973.2015.1065216

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. doi:10.1002/sim.8086

Petit-Bois, M., Baek, E. K., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2015). The consequences of modeling autocorrelation when synthesizing single-case studies using a three-level model. *Behavior Research Methods*. doi:10.3758/s13428-015-0612-1

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

Pustejovsky, J. E. (2018). Procedural sensitivities of effect sizes for single-case designs with

directly observed behavioral outcome measures. *Psychological Methods*. doi:10.1037/met0000179

Pustejovsky, J. E., & Ferron, J. (2017). Research Synthesis and Meta-Analysis of Single-Case Designs. In J. M. Kaufmann, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of Special Education* (2nd Edition., p. 63). New York, NY: Routledge.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*(5), 368–393. doi:10.3102/1076998614547577

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

Radley, K. C., O'Handley, R. D., & Labrot, Z. C. (2015). A comparison of momentary time sampling and partial-interval recording for assessment of effects of social skills training. *Psychology in the Schools*, *52*(4), 363–378. doi:10.1002/pits.21829

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554. doi:10.1111/j.1467-9876.2005.00510.x

Rindskopf, D. M. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*, *52*(2), 179–189. doi:10.1016/j.jsp.2013.12.003

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

Rohatgi, A. (2015, October). Webplotdigitizer. Zenodo. doi:10.5281/zenodo.32375

Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, *23*(2), 139–146. doi:10.1177/0963721414524773

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013a). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, 1–43.

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

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2013b). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*. doi:10.1016/j.jsp.2013.11.004

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, *6*(4), 228–237.

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

Swan, D. M., & Pustejovsky, J. E. (2018). A gradual effects model for single-case designs. *Multivariate Behavioral Research*, *53*(4), 574–593. doi:10.1080/00273171.2018.1466681

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case

designs: Baseline corrected Tau. *Behavior Modification*, *41*(4), 427–467. doi:10.1177/0145445516676750

Taylor, B. A., & Hoch, H. (2008). Teaching children with autism to respond to and initiate bids for joint attention. *Journal of Applied Behavior Analysis*, *41*(3), 377–391. doi:10.1901/jaba.2008.41-377

Thorne, S. J. (2005). *The effects of a group contingency intervention on academic engagement and problem behavior reduction in elementary school classrooms with at-risk students.* (Doctoral Dissertation). University of Kansas.

Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. doi:10.1080/17489530802505362

Verschuur, R., Didden, R., Lang, R., Sigafoos, J., & Huskens, B. (2014). Pivotal response treatment for children with autism spectrum disorders: A systematic review. *Review Journal of Autism and Developmental Disorders*, *1*(1), 34–61. doi:10.1007/s40489-013-0008-z

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wilke, C. O. (2018). *Ggridges: Ridgeline plots in 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggridges

Wilke, C. O. (2019). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=cowplot

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

Yoder, P. J., Ledford, J. R., Harbison, A. L., & Tapp, J. T. (2018). Partial-interval estimation of count: Uncorrected and poisson-corrected error levels. *Journal of Early Intervention*, *40*(1), 39–51. doi:10.1177/1053815117748407

Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259–3270. doi:10.1016/j.csda.2008.11.033

Appendix

R and package versions

R version 3.6.0 (2019-04-26)

Platform: x86_64-w64-mingw32/x64 (64-bit)

locale: *LC_COLLATE=English_United States.1252*, *LC_CTYPE=English_United States.1252*, *LC_MONETARY=English_United States.1252*, *LC_NUMERIC=C* and *LC_TIME=English_United States.1252*

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *pander(v.0.6.3)*, *ARPobservation(v.1.2.0)*, *ggridges(v.0.5.1)*, *cowplot(v.0.9.4)*, *colorspace(v.1.4-1)*, *kableExtra(v.1.1.0)*, *xtable(v.1.8-4)*, *readxl(v.1.3.1)*, *forcats(v.0.4.0)*, *stringr(v.1.4.0)*, *dplyr(v.0.8.1)*, *purrr(v.0.3.2)*, *readr(v.1.3.1)*, *tidyr(v.0.8.3)*, *tibble(v.2.1.3)*, *ggplot2(v.3.1.1)*, *tidyverse(v.1.2.1)*, *papaja(v.0.1.0.9842)* and *knitr(v.1.23)*

loaded via a namespace (and not attached): *tidyselect(v.0.2.5)*, *xfun(v.0.7)*, *haven(v.2.1.0)*, *lattice(v.0.20-38)*, *generics(v.0.0.2)*, *viridisLite(v.0.3.0)*, *htmltools(v.0.3.6)*, *yaml(v.2.2.0)*, *base64enc(v.0.1-3)*, *rlang(v.0.3.4)*, *pillar(v.1.4.1)*, *glue(v.1.3.1)*, *withr(v.2.1.2)*, *modelr(v.0.1.4)*, *plyr(v.1.8.4)*, *munsell(v.0.5.0)*, *gtable(v.0.3.0)*, *cellranger(v.1.1.0)*, *rvest(v.0.3.4)*, *codetools(v.0.2-16)*, *evaluate(v.0.14)*, *labeling(v.0.3)*, *highr(v.0.8)*, *broom(v.0.5.2)*, *Rcpp(v.1.0.1)*, *scales(v.1.0.0)*, *backports(v.1.1.4)*, *webshot(v.0.5.1)*, *jsonlite(v.1.6)*, *hms(v.0.4.2)*, *digest(v.0.6.19)*, *stringi(v.1.4.3)*, *bookdown(v.0.11)*, *grid(v.3.6.0)*, *cli(v.1.1.0)*, *tools(v.3.6.0)*, *magrittr(v.1.5)*, *lazyeval(v.0.2.2)*, *crayon(v.1.3.4)*, *pkgconfig(v.2.0.2)*, *xml2(v.1.2.0)*, *lubridate(v.1.7.4)*, *assertthat(v.0.2.1)*, *rmarkdown(v.1.13)*, *httr(v.1.4.0)*, *rstudioapi(v.0.10)*, *R6(v.2.4.0)*, *nlme(v.3.1-139)* and *compiler(v.3.6.0)*