The impact of response-guided designs on count outcomes in single-case experimental design baselines

Daniel M. Swan

University of Oregon

James E. Pustejovsky and S. Natasha Beretvas

The University of Texas at Austin

Author Note

Daniel M. Swan, Prevention Science Institute, University of Oregon

James E. Pustejovsky and S. Natasha Beretvas, Department of Educational Psychology, The University of Texas at Austin

Correspondence concerning this article should be addressed to Daniel M. Swan, e-mail: dmswan@uoregon.edu

A previous version of this work was presented at the American Educational Research Association annual convention, April 7, 2019 in Toronto, Ontario.

Abstract

In single-case experimental design (SCED) research, researchers often choose when to start treatment based on whether the baseline data collected so far are stable, using what is called a response-guided design. There is evidence that response-guided designs are common, and researchers have described a variety of criteria for assessing stability. With many of these criteria, making judgments about stability could yield data with limited variability, which may have consequences for statistical inference and effect size estimates. However, little research has examined the impact of response-guided design on the resulting data. Drawing on both applied and methodological research, we describe several algorithms as models for response-guided design. We use simulation methods to assess how using a response-guided design impacts the baseline data pattern. The simulations generate baseline data in the form of frequency counts, a common type of outcome in SCEDs. Most of the response-guided algorithms we identified lead to baselines with approximately unbiased mean levels, but nearly all of them lead to underestimates in the baseline variance. We discuss implications for the use of response-guided designs in practice and for the plausibility of specific algorithms as representations of actual research practice.

Keywords: single-case experimental designs, response-guided designs, behavioral observation

The impact of response-guided designs on count outcomes in single-case experimental design baselines

Single-case experimental designs (SCEDs) are an important tool for evaluating interventions developed for use with low-incidence populations, and are applied in fields such as communication disorders, rehabilitation, special education, and clinical and school psychology. Several types of SCEDs have been described (Ledford & Gast, 2018), all of which have in common the repeated measurement of an outcome over time on an individual case. In many types of SCEDs, measurements are organized in phases, beginning with a baseline phase to establish an initial pattern of responding, followed by a phase where the intervention is introduced. In multiple-baseline designs, the baseline and treatment phases are replicated across multiple cases, each beginning intervention at a different point in time, while in treatment reversal designs, the intervention phase is followed by a return-to-baseline and re-introduction of intervention(Gast, Ledford, & Severini, 2018; Gast, Lloyd, & Ledford, 2018).

In implementing an SCED involving baseline and intervention phases, the researcher must determine when to begin the intervention phase. Although methodologists have argued for using randomization procedures to determine phase change points (Kratochwill & Levin, 2010; Todman & Dugard, 1999), applied single-case researchers have questioned the value of randomized designs, and the approach remains uncommon in practice (Ledford, 2018; Wolery, 2013). More commonly, researchers graph the outcome data after every measurement occasion and determine when to change phases based on inspecting the data pattern—an approach termed *response-guided experimentation* (for the origin of this terminology, see Edgington, 1983).

A response-guided design involves making inferences about certain features of the data (Gast, 2014). These inferences are not summary inferences about the effectiveness of an intervention, but rather are judgments about patterns in the data that drive decisions about the timing of implementing an intervention. Typically, the analyst is concerned with “stability” in the current phase for any or all participants, where stability means that the baseline data are (a) not too variable and (b) not trending in the expected direction of the effect (Barton, Lloyd, Spriggs, & Gast, 2018; Joo, Ferron, Beretvas, Moeyaert, & Van den Noortgate, 2017; Kazdin, 1982; Kratochwill, Levin, Horner, & Swodoba, 2014). Ensuring stability is desirable for visual analysis because low variability makes it easier to discern whether an intervention has an impact on the average level of an outcome. Ensuring that there is no trend in baseline is desirable because a baseline trend (particularly a trend in the direction of therapeutic improvement) makes it difficult to visually distinguish between a projected natural change and change due to introducing the intervention.

Although visual inspection is the traditional method of drawing summary inferences from SCEDs, the past several decades have seen growing interest in other methods of analysis—particularly methods for meta-analysis of SCEDs (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Van den Noortgate & Onghena, 2003). Pustejovsky and Ferron (2017) described a number of benefits to the meta-analysis of SCED data. First, meta-analysis can provide a basis for drawing broader generalizations about intervention impacts than what can be supported by individual SCED studies, which are usually small and case-specific. Second, meta-analysis provides a way to measure the consistency or heterogeneity of intervention effects and to examine moderating factors that could explain why treatments work in some cases but not others. Third, meta-analyses can also provide insights into methodological practices, either by bringing to light issues in empirical findings or by examining how well the available research conforms to methodological standards and best practices.

Alongside interest in meta-analysis of SCED data, there has also been an interest in the application of statistical methods to individual cases or studies. Statistical methods for individual studies can provide a complement to visual analysis methods because the latter method does not provide a clear means of estimating the magnitude of the effect, but only of assessing whether or not a change in the outcome is related to the intervention (Kratochwill et al., 2014).

Despite recent interest in statistical analysis and meta-analysis of SCED data and a rapidly expanding array of new methods (Manolov & Moeyaert, 2017), there has been little consideration of how response-guided design practices might impact results of statistical analysis. All of the existing statistical methods that we are aware of treat the length of baseline and intervention phases as fixed aspects of the study's design, yet response-guided design practices would suggest that phase lengths are determined dynamically from the data pattern. Furthermore, response-guided designs might lead to restricted variability in the outcome data, with consequences for estimates and inferences from statistical analysis. Specifically, if the variability of the sample data is restricted with respect to the population process, the use of a response-guided design might affect the properties of effect size estimates that use the sample standard deviation, as well as standard errors and confidence intervals that characterize the degree of uncertainty in estimates.

Studying the impact of response-guided designs on SCEDs is difficult because we need precise operational definitions of how researchers make decisions about when to transition between phases. Researchers who use response-guided design practices typically do not precisely describe how they make phase change decisions, and in fact may not even report whether they used response-guided design practices at all. In a review of 101 multiple-baseline SCEDs published between 1998 and 2001, Ferron and Jones (2006) found that as many as 80% of the studies may have been response-guided, but only 31% of the studies were *explicitly* described as response-guided.

However, there is some guidance in applied textbooks (e.g., Gast & Spriggs, 2014; Kazdin, 1982) and the scant methodological research on the impact of response-guided designs on SCEDs (Ferron, Joo, & Levin, 2017; Joo et al., 2017) that we can use to make some educated guesses about appropriate operational definitions for response-guided designs. We propose the use Monte Carlo simulations to mimic some of the response-guided design practices used by researchers and thereby to assess their impact.

Monte Carlo simulations

One of the main available tools for studying the properties of statistical methods is Monte Carlo simulation. In a Monte Carlo simulation, the researcher generates many random samples of artificial data, based on a fully specified statistical model for the population. The researcher then applies a statistical method to each of the samples and compares the results to the known true parameters of the population model (in contrast, when applying statistical methods to real data, the absolute truth of the population is never known with certainty). Methodologists often use Monte Carlo simulations as part of developing and validating new statistical estimation methods or in evaluating competing methods (Morris, White, & Crowther, 2019). The methods are especially useful for studying the properties of statistical methods when their assumptions are violated—a situation where it is difficult to derive properties using theory alone. Response-guided designs may represent subtle violations of the assumption of independence or the assumed distributions of the outcomes, and so the use of simulation methods is an appropriate way to explore their impacts.

In order for simulations to provide meaningful findings regarding the properties of a method (such as response-guided design), the model underlying the simulations must be credible. Many simulations in educational and social science research assume that data are normally distributed. However, outcomes in real SCEDs are frequently reported as counts, rates of behavior, or proportions of time (Pustejovsky, Swan, & English, 2019; Shadish & Sullivan, 2011). Typical models for counts or proportions (such as the Poisson or binomial distributions) involve a connection between the mean level and variability of the outcomes, whereas models based on normal distributions usually do not involve mean-variance relationships. Because the variance of the outcomes is a key feature of many response-guided design approaches, it is important for simulations of such practices to consider models that involve different mean-variance relationships.

Study Aims

In this study, we draw on both methodological and applied texts as guidance for designing algorithms to mimic response-guided methods. Using Monte Carlo simulation, we examine how the use of these hypothetical methods affects the features of SCED data. The algorithms we designed use two sets of criteria drawn from Kazdin (1982) (a classic and influential practitioner text), four sets of criteria from Gast and Spriggs (2014) (a modern practitioner text), and the virtual visual analyst developed by Ferron et al. (2017) and Joo et al. (2017). To our knowledge, Ferron et al. (2017) and Joo et al. (2017) are the only previous studies to examine response-guided designs in the context of statistical models. These studies proposed and studied one method for simulating a response-guided design using normally-distributed data. The present study builds on Ferron et al. and Joo et al. by examining multiple potential response-guided design algorithms and by simulating count outcome data. The first contribution is important because, in practice, it is likely that researchers use an array of response-guided design practices other than the virtual visual analyst algorithm proposed in Ferron et al. (2017). The second contribution is important because normal distributions may not be a good representation of the types of data often collected in SCEDs.

The present study is motivated by two research questions:

- What are the distributions of stable baseline phase lengths from different response-guided algorithms?

In practice, SCEDs often involve short baseline phases. A review by Pustejovsky and colleagues (2019) found a mean baseline length of 11 observations, with a median of 7, indicative of a distribution made up of short baselines and a small number of much longer baselines. If the baselines that result from a particular algorithm are generally very long, this suggests that the algorithm might be unrealistic for practical use. Researchers are unlikely to apply criteria that never allow them to move to intervention phases. Additionally, algorithms for which all baselines reach stability are potentially not very discriminating in terms of the "stability" of the data, although it is likely that some researchers prefer rules that are minimally discriminating. Information about the average baseline lengths, the proportion of baselines that are found stable in reasonable number of observations, and combined discriminating power may also be useful to applied researchers for deciding on the sort of criteria they may want to use in their response-guided designs.

- To what degree does the use of response-guided algorithms create systematic bias in the mean level or variability of sample baselines?

SCED researchers may be interested in using response-guided design practices that produce data which are amenable to statistical analysis in addition to traditional visual analyses, and researchers interested in the secondary analysis of SCED data will want unbiased estimates of treatment effects. It is therefore important to know whether using response-guided design practices creates systematic biases. Biases in the mean level can impact treatment effect estimates, and biases in the variability can impact standardized treatment effect estimates, standard errors, as well as the weights for effect sizes in meta-analysis.

There are a variety of SCED designs (treatment reversals, multiple baselines, etc.) and potential post-treatment trajectories (immediate treatment effects, linear trends, non-linear trends) that could be studied through simulation, and each of these design and post-treatment trajectory combinations represent a potential study on their own. In this study, we focus on simulated count data from SCED baselines without any time trend (i.e., no systematic pattern of decreasing or increasing values). This simple scenario is a useful starting point because it presents favorable conditions for conducting an SCED and using statistical analysis. Thus, this scenario can provide some insight into the impact of response-guided designs on the results of statistical methods applied to SCEDs. We should note, however, that the scenarios we examine all assume that the true data generating process has a stable level. Consequently, we do not examine the performance of the algorithms in detecting baseline instability, leaving such questions for further research.

Methods

In what follows, the response-guided algorithms and data generating models are described in greater detail, the conditions under which we generated the simulated SCED baselines are explained, and the performance criteria we used to evaluate the simulations are described.

Response-guided algorithms

Kazdin (1982) noted that there was no agreed-upon method for determining stability in SCEDs. He identified some empirical examples to demonstrate possible methods for determining stability, and we followed these examples to define two response-guided algorithms. Drawing on the example of Wilson, Robertson, Herlong, and Haynes (1979), we considered a series stable if the last three observations in the series were within ±5% of the mean of the full series. We called this the Kazdin 10% rule, abbreviated as **Kaz10**. Drawing on the example of Scott et al. (1973), we considered a series stable if the last three observations in the series were within ±7.5% of the mean of the full series. We called this the Kazdin 15% rule, abbreviated as **Kaz15**.

We based our next set of algorithms on Gast and Spriggs (2014). Between the time that we performed this research and the results were published, a newer version of this chapter (Barton et al., 2018) was published in Ledford and Gast (2018). Any updates to the text are not accounted for in our algorithms, although we intend that our future research in this area will take any substantial changes into consideration.

Gast and Spriggs (2014) described a number of ways to characterize stability in SCEDs. Although not explicitly framed as rules for transitioning between phases, we used their descriptions as guidelines for other ways a researcher might decide a baseline is stable. They describe four methods of detecting “level” stability, or stability when there is no trend in the data, all of which are defined in terms of a horizontal envelope or bandwidth. The authors describe bandwidths as large as 25% of the median being appropriate with a small amount of data, and bandwidths as small as 10% of the median being appropriate for a large amount of data. To make this method fully operational, we used examples described by Gast and Spriggs (2014) as guidance. Let w denote the width of the envelope used to determine stability. We assume that the stability envelope width w is determined by the number of observations in baseline up to that point, denoted as n_b, and by the median of the baseline data up to that point, denoted $\tilde{m}_b$. Specifically, we set w equal to 25% of $\tilde{m}_b$ when n_b is less than or equal to 5 and equal to 10% of $\tilde{m}_b$ when n_b is greater than or equal to 20; for intermediate values, we assume that w decreases linearly by 1% of m_b for each additional observation between 5 and 20. Figure 1 displays the stability envelope visually.

For the first method, full-series baseline stability (**GSFull**), we considered a baseline stable when at least 80% of the observations in a series are within a horizontal envelope (called the ”80%/20%” rule in the text). For the following three rules, no explicit guidance was provided by the authors on what was considered an appropriate width, so we approximated appropriate widths based on the spread implied by the 80%/20% rule. For the second method (**GSFinal**), we considered a baseline stable if all of the final three observations are within a horizontal envelope of width w. For the third method, termed “absolute stability” by the authors (**GSAbs**), we considered a baseline stable if the absolute value of the difference between the first and last data point was less than $2w$. For the fourth method, termed “relative stability” by the authors (**GSRel**), we considered a series stable if the absolute difference between the median of the data from the first half of the baseline and the median of the data from the second half of the baseline (excluding the middle point for an odd number of observations) was less than $2w$.

Gast and Spriggs (2014) also advised assessing trend using a freehand or split-middle line, using the same bandwidth as the 80%/20% rule around the line. We focused on the split-middle line because it can readily be automated. In conjunction with the level stability criteria described above, we assumed that a baseline series would be considered stable if the slope of the split-middle line is contra-therapeutic and 80% or more of the points in the baseline are within an envelope of a width w around the split-middle line. Drawing on general advice from Gast and Spriggs (2014), we also required that any therapeutic trend must be smaller than $1.5w/n_b$. This ensured that small level changes that would be considered acceptable to visual analysts would not cause the algorithm to keep collecting data, while also ensuring that a baseline would not end with a therapeutic trend.

For the final method, we drew on the methodological literature and implemented the virtual visual analyst for baseline stability (**VVA**), as described in Joo et al. (2017). The authors developed the virtual visual analyst to mimic human judgment about whether and when to extend a phase within a single-case experimental design, so that they could study the performance of inferential procedures such as masked visual analysis using Monte Carlo simulation. They provided precise operational definitions, allowing us to match their proposed algorithm exactly. This version of the algorithm assumes that the therapeutic direction of treatment is an increase in the behavior. Let s_b denote the standard deviation of the baseline observations collected thus far. We considered the baseline stable if all of the following conditions were met:

1. the ordinary least squares (OLS) regression slope of baseline observations is less than $0.5s_b$,

2. the OLS slope of the final three baseline observations is less than $0.5s_b$,

3. the difference between the last observation in the baseline and the mean of the

entire baseline is less than $2s_b$, and

4. the difference between the mean of the last half of the baseline observations and the mean of the first half of the baseline observations is less than $1.5s_b$.

Careful examination of the algorithms suggests that response-guided design practices from practitioner texts are generally geared toward assessing the change in level, degree of variability, and sometimes trend in the direction of the expected effect. The stability rules from Gast and Spriggs (2014) appear to be refinements on the ones described by Kazdin (1982). Because they are defined in terms of the median rather than the mean, the algorithms from Gast and Spriggs (2014) will not be affected by outlying values to the same degree as those from Kazdin (1982). The algorithms from practitioner texts come at the issue of stability from the perspective of a visual analyst, and therefore are defined in terms that can easily be calculated from plotted data. The VVA comes at stability from the perspective of a quantitative analyst, who assumes that the spread in the sample is representative of the generating process, and who is concerned with avoiding therapeutic trends in the baseline and large changes in level between the beginning and end of the baseline.

Finally, these rules for stability are likely only useful in contexts where the outcome measure has an absolute zero. Measures such as counts (which our study focuses on) or rates have meaningful zeros, but other measures used in SCEDs may not.

Data Generating Models

To investigate the consequences of response-guided designs for baseline data patterns, we simulated artificial baseline phase data based on three distinct data generating models: an auto-correlated Poisson distribution, a gamma point process, and a normal (Gaussian) distribution.

Auto-correlation, where the value of a given observation depends upon one or more of the previous observations, is a common concern in SCED data (Shadish & Sullivan, 2011), and adequately accounting for it in statistical models is an ongoing area of research in SCED methods (Shadish, 2014). In light of these concerns, we first examined a model that generated auto-correlated observations: an auto-correlated Poisson distribution. This model generates count outcomes, similar to the type of data that would be generated using a frequency counting procedure for systematic direct observation of a behavior. Taken individually, each observation follows a Poisson distribution, which is a common model for count outcomes (Fox, 2008). The model introduces auto-correlation using a method called binomial thinning (McKenzie, 1988). Binomial thinning generates dependent observations by summing an independent Poisson draw with a draw from a binomial distribution, where the parameters of the binomial distribution are determined by previous observations. We assumed that the data followed a first-order auto-regressive structure, where the correlation between an observation at time s and an observation at time t is $\rho^{|t-s|}$.

The second model was a gamma point process (Rogosa & Ghandour, 1991), which also produces count outcomes. Briefly, the gamma point process is a model for the number of behaviors ("points") observed in a specified time period, where each instance of behavior is instantaneous and the waiting times between behaviors follow a gamma distribution. We used this model in order to examine the consequences of different degrees of variability (or dispersion) in the outcome data, in the absence of auto-correlation. Whereas Poisson-distributed data have a variance that is exactly equal to its mean, data produced by the gamma point process have a variance that is approximately proportional, but not equal, to its mean. With a gamma point process, the variance of an observation may be larger than its mean (over-dispersion) or smaller than its mean (under-dispersion). This is roughly equivalent to what is called the quasi-Poisson mean-variance relationship used in generalized linear models (McCullagh & Nelder, 1989). The problem of over- or under-dispersed data is common enough that some texts suggest using the quasi-Poisson mean-variance relationship as a matter of course when modeling count outcomes (Fox, 2008). The inspection of count data from SCEDs has also shown evidence of over- and under-dispersion (Pustejovsky et al., 2019). In the simulations, we generate baseline data with a specified mean, μ, and relative dispersion $\kappa \neq 0$, so that the variance of each observation is $\kappa\mu$. Values of κ less than 1 correspond to under-dispersion, while κ larger than 1 corresponds to over-dispersion. When $\kappa = 1$, the gamma point process is equivalent to a Poisson distribution (without auto-correlation).

The third and final model was a traditional normal distribution, with and without autocorrelation. We generated normally-distributed data in order to look for discrepancies between normally-distributed model and the other count models, and to link our results to the existing methodological literature that has largely focused on the normal distributions for generating outcome data.

Simulation Conditions

We conducted the simulations using the R statistical computing environment (R Core Team, 2020). Table 1 displays the simulation conditions for each of the data generating processes. For each data generating process, all conditions were crossed. Because the Poisson closely approximates a normal distribution when the counts are high-incidence (Johnson, Kemp, & Kotz, 2005), we focused on counts with lower incidence, with mean levels no greater than 25. For the binomial thinning model, 0.2 is the mean level of autocorrelation found by Shadish and Sullivan (2011) in their review, and 0.4 is one standard deviation above the mean. We created functions to generate data according to the binomial thinning model.

For the gamma point process, we selected dispersion levels based on an empirical analysis of baseline data from seven published systematic reviews of SCEDs (Pustejovsky et al., 2019). We omitted the case with $kappa = 1$ because it is mathematically equivalent to a Poisson distribution with zero auto-correlation, and so is redundant with the first simulation. We used the ARPobservation package

(Pustejovsky, 2018) to generate this data.

We generated normally-distributed data with a mean of 5, with variances of 0.25, 1, and 2.25 (corresponding to standard deviations of 0.5, 1, and 1.5). We chose a mean of 5 for two reasons. The first reason is that we use relative bias as a performance criteria, and we needed to use a value other than zero for the mean baseline in order to avoid dividing by zero. Additionally, we wanted to avoid simulating observations less than zero, which is clearly not possible for counts. The autocorrelation values were chosen using the same logic as the binomial thinning simulation conditions. We used the R function ‘arima.sim‘ to generate normally-distributed data with first-order auto-correlation, and the R function ‘rnorm‘ to generate uncorrelated normally-distributed data.

For each data generating model and each combination of conditions, we simulated data from 5000 baseline phases, each 100 observations in length. We then analyzed the stability of each simulated baseline data series. Specifically, for each of the algorithms that we have described, we found the minimum phase length where the data series met stability criteria, up to a maximum length of 100 observations. Based on the reported characteristics of empirical SCED data (Pustejovsky et al., 2019), baselines longer than 100 observations are very unlikely in real data. Within each set of conditions and algorithm, we tracked the total number of observations that were found stable at any point between 3 and 100 observations, inclusive.

For the subset of simulated baseline data series that met stability criteria, we calculated the relative bias of the baseline level and baseline variance for each combination of data generating model, conditions, and algorithm. If the true value of a parameter is θ and its estimate is $\hat{\theta}$, relative bias is defined as $\mathrm{E}\left(\hat{\theta} - \theta\right)/\theta$. Following the guidelines suggested by (Hoogland & Boomsma, 1998), we indicate sets of conditions that exceed ±5% of the relative bias, because biases of that magnitude may be of particular concern. In total, this simulation had $3 \times 3 + 3 \times 4 + 3 = 24$ sets of conditions. Code for replicating all of the simulations is available (Swan, Pustejovsky, & Beretvas, 2020).

Results

We present the results of the simulation regarding, in turn, the distributions of the baselines, the bias of the baseline mean, and the bias of the baseline variance. Although methodological studies typically describe any biases before considering other performance criteria, we believe that the typical phase lengths which are the result of using these algorithms may be of the most interest to applied SCED researchers.

Distributions of the baselines

For this section, we focused primarily on the baselines that reached stability within the first 20 observations. This range provided the clearest evidence of the differences between algorithms, and we believe that most applied researchers are likely to be interested in baselines of fewer than 20 observations. For most of the response-guided design algorithms, any baseline that would eventually be considered stable was usually stable prior to 20 observations, except for the two Kazdin algorithms.

Figure 2 displays the cumulative percentage of stable baselines at a given observation length up to 20 observations for the independent Poisson data. The value at $x = 3$ is the percentage of stable baselines at three observations, the value at $x = 4$ is the percentage of stable baselines at three observations plus the percentage of stable baselines at four observations, and so on. Generally speaking, higher values of the mean led to larger percentages of the baselines that were eventually considered stable. When the generating baseline mean was low (a mean of five observations per session), relatively few data series achieved stability according to the Kaz10, Kaz15, or GSFull algorithms—between 10% and 20% of baselines by 20 observations. As the generating mean increased, the fraction of data series that achieved stability by 20 observations increased, up to around 20% for Kaz10, around 40% for Kaz15, and around 60% for the GSFull algorithm. The GSFinal found a larger number of baselines stable by 20 observations, about 50% when $\mu = 5$ and 95% when $\mu = 25$. The increasing trend observed in the Kaz10, Kaz15, and GSFinal algorithms continued all the way up to 100 observations, with all or nearly all of the baselines eventually reaching stability in the case of the GSFinal algorithm. The cumulative percentage of baselines found stable by the GSFull algorithm was essentially unchanged after 20 observations, with most baseline ending at three observations (the first possible observation for stability) or five observations (the first time a baseline with a contra-therapeutic trend can be found stable). The GSAbs and GSRel algorithms found nearly all of the baselines stable by 20 observations, with the rest achieving stability by 30 observations. The VVA algorithm found all of the baselines stable by 20 observations.

The relationship between the mean level and stability is likely a consequence of the fact that for the Poisson and gamma point process data, the overall spread of the data (the standard deviation) increases proportionally to the square root of the mean. As the mean increases, the rules based on a bandwidth that is proportional to the mean represent less restrictive conditions for the data to meet.

Figure 3 displays the cumulative percentage of stable baselines at a given observation length for a generating mean of $\mu = 15$ across varying degrees of autocorrelation. Increasing autocorrelation led to more baselines being found stable, and those baselines that were stable were also shorter in length. The only exception was the VVA algorithm, where increasing autocorrelation led to slightly longer baselines.

For the gamma point process model, the influence of different values of the true mean followed the same pattern as with the auto-correlated Poisson model. We therefore focused on the influence of varying degrees of dispersion. Figure 4 displays the cumulative percentage of stable baselines at a given observation length for gamma point process outcomes with a mean of $\mu = 15$ across varying degrees of dispersion. For comparison, we have also included the results from the Poisson distribution (i.e., with unit dispersion) and zero auto-correlation.

For the Kaz10, Kaz15, and GSFull algorithms, the degree of dispersion strongly affected the chance that the data series achieved stability. For instance, when the generating mean was $\mu = 15$ and the overdispersion $= 2.5$, only about 5% of baselines were considered stable by 20 observations by the Kaz10 algorithm, whereas more than 30% of the baselines in the heavily underdispersed case (dispersion $= 0.4$) were considered stable at the same mean by the same algorithm. In general, outcomes with lower variability were more likely to be seen as stable, when holding the value for the mean constant. The only exception was the VVA, where the influence of over- and under-dispersion was not as strong or consistent. As with the Poisson distribution, data series simulated from a gamma point process always achieved stability well before 100 observations according to the GSAbs, GSRel, and VVA algorithms.

Figure 5 displays the cumulative percentage of stable baselines at a given observation length for normally-distributed baselines across varying levels of autocorrelation when $\sigma^2 = 1$. While autocorrelation does have a small impact on the proportion of cases that reach stability for the Kaz10, Kaz15, GSFull, and GSFinal algorithms, the differences were only noticeable in the case of the GSFull algorithm, where the relationship between autocorrelation and stability was consistent with the Poisson results.

Figure 6 displays the cumulative percentage of stable baselines at a given observation length for normally-distributed baselines across varying levels of the variance when the data are independently distributed. The effect of increasing variance while holding the mean stable is similar to the impact of the increasing mean in both of the Poisson cases. As the variability of the data increases, the percentage of baselines that ever reach stability decreases (with the exception of the VVA). This is unsurprising, all of the methods other than the VVA are based on the bandwidth around the mean or median, so reducing the variability of the data will increase the number of stable baselines.

One notable aspect of these results is that, across any of the conditions examined, the GSFull algorithm never determined that a baseline was stable after five observations. This may be an indication that the GSFull algorithm is overly restrictive, possibly due to a combination of a bandwidth that is too narrow and an algorithm that considers all of the observations in the series. In contrast, several of the other algorithms consider only a subset of the observations (e.g., the most recent three observations) when determining stability.

Bias of the sample mean

As we noted previously, biases in the baseline sample mean have potential to bias treatment effect estimates from statistical models. Thus, both applied researchers and researchers interested in the secondary analysis of SCED data should be aware of the potential consequences of using response-guided designs for estimation of baseline mean levels.

Figure 7 displays the relative bias of the baseline mean for the Poisson-distributed outcomes that reached stability within 100 observations. The relative bias of the baseline mean was less than 3% for most of the algorithms and under most conditions. The one exception was the GSFull algorithm, where the sample mean had a relative bias of 5% or more when the generating mean was small ($\mu = 5$) and there was positive autocorrelation in the errors. The GSFinal algorithm also led to small biases (of less than 3%) when the generating mean was small ($\mu = 5$). The GSFull algorithm is more restrictive than other algorithms, considering each observation as a part of stability, as opposed to a subset of the observations. It is likely that only baselines with a large sample mean (and consequently larger bandwidth) with respect to the generating conditions were able to meet the criteria of the algorithm. While the GSFinal algorithm was not notably more restrictive than the Kaz10 or Kaz15 algorithms, it did find a larger number of very short baselines stable. Similar to the GSFull algorithm, these shorter baselines likely had larger means to meet the criteria of the algorithm.

Figure 8 displays the relative bias of the sample mean for stable data series simulated from the gamma point process model that reached stability within 100 observations. As with the Poisson model, baseline biases were generally less than 3%, although there were two exceptions. First, the Kaz10 algorithm caused negative bias of approximately 5% when the generating mean was small ($\mu = 5$) and the data were highly overdispersed (2.5). In the case of overdispersed data with a small mean, these baselines likely had observations with very low counts that the algorithm found stable. Second, the GSFull algorithm created positive bias when the mean was small ($\mu = 5$) and there was any overdispersion, in which case the baseline mean was overestimated by as much as 10%. This behavior was likely due to the same requirement for a large sample mean in order for the bandwidth to be large enough for the baseline to meet the criteria of the algorithm.

Figure 9 displays the relative bias of the baseline mean for normally-distributed baselines that reached stability within 100 observations. The biases here were relatively small, except for the combination of the GSFull algorithm and the largest value of the variance ($\sigma^2 = 2.25$), likely for the same reasons as noted in the Poisson-distributed baselines. In nearly all of these cases, the variance of the outcome with respect to the generating mean was much smaller than in the Poisson or gamma point process case, which may explain why there were so few cases with noticeable bias of the baseline means.

Bias of the sample variance

Perhaps even more important than bias in the baseline sample mean is a bias in the baseline sample variance. Biases in the sample variance have the potential to impact standard errors and weights for regression-based models, as well as the magnitude of any effect size that is standardized by the sample variance, such as the within-case standardized mean difference.

Figure 10 displays the relative bias of the baseline variance for Poisson-distributed outcomes that reached stability within 100 observations. In nearly all cases, using a response-guided algorithm led to an underestimate in the variance of the baselines. The Kaz10, Kaz15, and GSFinal algorithms averaged around a 20% underestimate of the variance, and the GSFull algorithm averaged around a 60% underestimate of the variance. Due to the restrictive nature of the GSFull algorithm, it required baselines with low sample variances in addition to high sample means in order to meet the criteria of the algorithm. The GSAbs and GSRel algorithms produced less extreme biases, with underestimates in the 5-10% range when there is no autocorrelation in the errors.

Only one algorithm, the VVA, produced unbiased sample variances when there was no autocorrelation between the observations. This is likely because the VVA stability criteria are defined relative to the sample standard deviation, rather than the sample mean. It makes no judgments about the degree of variability, and only focuses on ensuring there is no therapeutic trend or sudden changes in level.

Across all algorithms, higher autocorrelation led to larger, negative biases in the sample variance. The connection between autocorrelation and bias of the sample variance arises for two reasons. The first is that variance estimates from small samples of autocorrelated data are negatively biased, and many of these baselines are quite short—less than ten observations. The second reason has to do with the spread of the variance estimates and their interaction with the algorithms that use a bandwidth. Let us denote the sample variance as S^2. When autocorrelation is present, there is more variability in the value S^2 across samples than when observations are independent. This means that samples with both larger and smaller variances are more likely than in the independent case. Early on, the baselines with small variances will be considered stable by algorithms using a bandwidth. Baselines with larger variances will not be considered stable by algorithms using a bandwidth and will continue to accrue observations. As the number of observations in a baseline increases, the sample variance will typically become less variable, so that the baseline will eventually be deemed stable by algorithms that use a bandwidth. Thus, for the algorithms defined in terms of a bandwidth, baselines with a very large sample variance are censored, which means that only samples with average variances or smaller-than-average variances are ever considered stable. As a consequence, the average variance of the stable baselines is smaller than the generating model.

Figure 11 displays the relative bias of the baseline sample variance for data series simulated from the gamma point process model that reached stability within 100 observations. Once again, the VVA algorithm produced data series with close-to-unbiased sample variances. The baselines found stable by the GSAbs and GSRel algorithms had variances that were underestimated by 5 to 10%. The Kaz10, Kaz15, and GSFinal algorithms led to underestimated variances of about 20%. For all these algorithms, the degree of bias was not strongly related to the level of dispersion used to generate the data series. In contrast, the GSFull algorithm led to large underestimated variance, ranging from 40 to 70%, with the largest underestimates occurring when the distribution was more over-dispersed, likely for the same reasons we discussed in the Poisson case.

Figure 12 displays the relative bias of the baseline variance for normally-distributed baselines that reached stability within 100 observations. Just as with the other models, the VVA had approximately unbiased variances when the data were independent. The GSAbs algorithm led to variances that were underestimated by about 5% when the observations were independent. In most other cases, the variances were underestimated, with negative biases of as much as 60% with the GSFull algorithm and the largest variance condition. Stronger autocorrelation led to larger negative bias of the sample variance, consistent with the pattern observed in the auto-correlated poisson model. Larger generating variances led to more pronounced, negative biases in the sample variance.

The one notable difference with the normal model is the Kaz10 and Kaz15 algorithms, where the smallest variance condition had the largest negative biases. When the data are independent, the differences are negligible. However, as autocorrelation increases, the relative bias becomes notably larger. This is likely related to the fact that negative biases of small magnitudes have large values when scaled by parameters of less than 1. For instance, a negative absolute bias of only .025 becomes a relative bias of 10% when the parameter it is being scaled by is 0.25. This, combined with the tendency of autocorrelation to produce underestimates of the variance for small samples lead to larger underestimates for those two algorithms when the variance was small and the autocorrelation larger.

Discussion

Looking across simulation conditions, the results indicated several general trends. First, higher generating means (for the Poisson conditions) or lower variances (for the normal-errors conditions) and larger autocorrelation led to shorter baselines and more baselines being found stable within 100 observations. In almost all cases, the estimated baseline mean was approximately unbiased for data series determined to be stable. In contrast, the sample variances were almost always underestimates of the true degree of variance in the baseline outcomes. The only algorithm whose behavior was inconsistent with these general patterns was the VVA.

The VVA works very differently than the other response-guided design algorithms. It simply uses the baseline sample standard deviation as a benchmark to ensure that there are no therapeutic trends in the data. The larger the sample standard deviation of the data, the *less* restrictive the VVA becomes, irrespective of the sample mean. The only reason that larger means led to a higher number of baselines being found stable was that, for several of the models, increasing the mean level also led to a corresponding increase in the standard deviation of the data. This also explains why increasing autocorrelation decreased the average length of the baselines for this algorithm. Increased autocorrelation caused a slight, negative bias in the variance when there were a small number of observations, which meant that the VVA was more restrictive in these cases.

Across all of the data generating models we examined, the Kaz10, Kaz15, and GSFull algorithms rarely allowed for a baseline reach stability. For Kaz10 and Kaz 15, the baselines that are stable have lengths that are relatively evenly distributed across the full range from 3 to 100 observations. In contrast, the baselines from Pustejovsky et al. (2019) look more like the data from simulated SCEDs in which the GSFinal, GSAbs, GSRel, and the VVA algorithms were applied, with most baselines ending within 5-10 observations and a long tail of a few long data series. When we consider how likely the algorithms are to have been used in practice, our conclusions are necessarily tentative. This is particularly true of the algorithms based on Gast and Spriggs (2014), as these were not explicitly described as response-guided criteria. If the assumptions underlying our data generating models are unrealistic, these results may be incorrect. Assuming that the data generating models are realistic, it seems unlikely that the Kaz10, Kaz15, and GSFull algorithms are being used in practice in the way that we have operationalized them. Although we cannot say with certainty that the GSFinal, GSAbs, GSRel, and the VVA algorithms represent actual practice, they do represent more plausible approximations for how response-guided design might be used in real research.

Implications for methodological research

Findings from this simulation study have several implications for further research on and development of methods for statistical analysis of SCED data. First and foremost, there is a need for a more detailed understanding of the stability criteria that are actually used in practice. To address this need, it would be useful to conduct systematic reviews, interviews, or surveys with applied researchers to learn about current research practices related to response-guided design.

Second, future methodological studies should consider the implications of response-guided designs for the properties of existing and new statistical methods. As we have noted, most existing statistical procedures for SCED data assume (perhaps implicitly) that phase lengths are fixed and unrelated to the data pattern during the baseline phase. Thus, there is a need to study whether use of response-guided designs may lead to bias in effect size estimation (Pustejovsky, 2019), multi-level modeling methods (such as those proposed by Van den Noortgate & Onghena, 2003), and statistical procedures that may be developed in the future. Similarly, non-overlap methods such as the percentage of non-overlapping data (Schlosser, Lee, & Wendt, 2008; Scruggs, Mastropieri, & Casto, 1987) or non-overlap of all pairs (Parker & Vannest, 2009) might be affected by the use of response-guided designs, due to their sensitivity to other operational details such as the number of observations in the baseline or treatment phases (Allison & Gorman, 1994).

Third, and more broadly, methodologists interested in SCED research should carefully consider their data generating model when simulating SCED data. The mean-variance relationship of our data generating models had important implications for stability. Although the normally-distributed data generally performed similarly to the Poisson and gamma point process data in the simulations that we presented, we had to set our generating conditions carefully in order to yield meaningful results.

Implications for applied research

Our simulation results suggest that use of certain response-guided designs may have consequences for the inferences drawn from SCED data. Further, not all response-guided algorithms behaved identically. The specific criteria used to assess stability and determine when to transition between phases mattered for the properties of estimates derived from the baseline data. Given the need to better understand how response-guided designs are used in practice, researchers should take care to report the precise algorithm or decision rule that they used in conducting an SCED. Describing their decision rules to the precise degree we have outlined in this paper might be difficult in practice. Textbook chapters such as Barton et al. (2018) provide an example for the level of detail needed to operationalize response-guided methods. To facilitate greater transparency and replicability, reviewers and editors of journals that publish SCED research should emphasize more detailed reporting of response-guided design practices.

As we have noted, many existing analysis methods for SCED data that are intended to complement or supplement visual analysis assume (perhaps implicitly) that phase lengths are fixed and unrelated to the data pattern during the baseline phase. Baselines with reduced variability may have underestimated standard errors with increased Type I error as a consequence, and biased weights in the context of meta-analysis. Researchers might consider specifying a suitable number of observations upon which to end the observation phase. With precise operational definitions, it might even be possible for methodologists and applied researchers to work together to create decisions rules around stability that help meet the goals of response-guided practices (i.e., aiding visual analysis) while also minimizing potential consequences such as the biases we have described.

Limitations and future directions

Moving forward, researchers might look outside education research for ways to accommodate response-guided designs in statistical models. One area of active development is adaptive clinical trials, which are used to evaluate medical interventions. Adaptive designs allow for adjusting the design of a trial, such as the probability that a new participant receives a novel intervention, while the study is underway, using data gathered over the study's course (Bhatt & Mehta, 2016). Methodological developments in this area may be relevant to the challenges of analyzing response-guided SCEDs in education and communication sciences and disorders.

The algorithms we have described and investigated in this simulation study represent an initial attempt to emulate the practice of single-case researchers who use response-guided designs. However, we have studied only a limited set of algorithms, and it is quite likely that we have not captured the full range of how SCED researchers assess baseline stability in practice. As primary researchers describe their methods more systematically and precisely, we can refine our algorithms and continue to study the impact of response-guided design practices.

A further limitation of this simulation study arises from looking only at selected data generating models. The results of this study are only meaningful insofar as the models that we have studied are reasonable approximations for the features of real data collected in practice. There are other potential data generating models for SCEDs, such as the negative binomial for counts, or the binomial and beta-binomial for proportions. These models exhibit different mean-variance relationships than the models we used in this study, and so the impact of response-guided methods on data generated from these alternative models might differ in important ways. Future research should examine the impact of additional data generating models with differing mean-variance relationships to understand how they interact with response-guided designs.

In addition to other mean-variance relationships, future research should examine data which are not temporally stable, where features of the data series such as the level or variability across time. As we noted in a previous section, this study is unable to address questions about the utility of these algorithms for diagnosing unstable data. By applying these algorithms to data series with systematic time trends, it would be possible to characterize how well these algorithms work to diagnose certain kinds of instability and to investigate the consequences of response-guided designs for other features of baseline data series, such as trend estimates.

Even based on the limited set of data generating models examined here, our findings indicate that there may be biases present in some SCED studies that used some form of response-guided practices. The fact that variance estimates from response-guided designs can be substantially biased has implications for treatment effect estimates commonly used by SCED researchers. Effect sizes such as the within-case standardized mean difference (Gingerich, 1984) or the non-overlap of all pairs (Parker &

Vannest, 2009) are defined in terms of the degree of variation in the baseline data series. Consequently, we suspect that estimates of these effect sizes may be inflated by response-guided design practices. Future work should investigate how response-guided design practices, applied to data from different data generating models, may be systematically biasing commonly used effect sizes for SCEDs.

A final limitation of this study is that we only considered algorithms for assessing stability of a single data series. In multiple-baseline designs, stability judgments may be made on the basis of multiple series simultaneously. Hybrid designs, such as a multiple-baseline with reversals, could complicate the impact of response-guided design practices even further. Likewise, if researchers collect more than one outcome measure in a study, then they will need to make stability judgments about one or more potentially correlated outcomes. Researchers' use of response-guided designs might also be affected by external constraints, such as school schedules. Researchers might face time constraints that require shifting criteria as the number of observations increases. They may have expectations about the impact of the interventions, which could influence the degree of variability or trend that is acceptable in the baseline phase. Future research should consider how the complex designs present in modern SCED studies (Moeyaert, Akhmedjanova, Ferron, Beretvas, & den Noortgate, 2020) and the response-guided practices used by applied researchers may interact to impact the features of SCED data and results of statistical analyses. Applied researchers can aid such investigations by carefully and specifically articulating the response-guided criteria that they use in practice.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*(6), 621-631.

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

Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford & D. L. Gast (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (p. 179-214). New York, NY: Routledge.

Bhatt, D. L., & Mehta, C. (2016). Adaptive designs for clinical trials. *New England Journal of Medicine*, *375*(1), 65–74.

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.

Edgington, E. S. (1983). Response-guided experimentation. *Psyccritiques*, *28*(1), 64-65.

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education*, *75*(1), 66-81.

Ferron, J. M., Joo, S.-H., & Levin, J. R. (2017). A Monte Carlo evaluation of masked visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis*, *50*(4), 701-716.

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

Gast, D. L. (2014). General factors in measurement and evaluation. In D. L. Gast & J. R. Ledford (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (p. 85-104). New York, NY: Routledge.

Gast, D. L., Ledford, J. R., & Severini, K. E. (2018). Withdrawal and reversal designs.

In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (p. 215-238). New York, NY: Routledge.

Gast, D. L., Lloyd, B. P., & Ledford, J. R. (2018). Multiple baseline and multiple probe designs. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (p. 239-281). New York, NY: Routledge.

Gast, D. L., & Spriggs, A. D. (2014). Visual analysis of graphic data. In D. L. Gast & J. R. Ledford (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (p. 176-210). New York, NY: Routledge.

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

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367.

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

Joo, S.-H., Ferron, J. M., Beretvas, S. N., Moeyaert, M., & Van den Noortgate, W. (2017). The impact of response-guided baseline phase extensions on treatment effect estimates. *Research in Developmental Disabilities*, *79*, 77-87. doi: https://doi.org/10.1016/j.ridd.2017.12.018

Kazdin, A. E. (1982). *Single-case research designs: methods for clinical and applied settings.* New York: Oxford University Press.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 124-44. doi: 10.1037/a0017736

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swodoba, 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* (p. 91-125). Washington, DC: American Psychological Association.

Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation*, *39*(1), 71-90. doi: 10.1177/1098214017723110

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences.* New York, NY: Routledge. doi: https://doi.org/10.4324/9781315150666

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 ed.) (No. 37). London ; New York: Chapman and Hall.

McKenzie, E. (1988). Some ARMA models for dependent sequences of poisson counts. *Advances in Applied Probability*, *20*(4), 822. doi: 10.2307/1427362

Moeyaert, M., Akhmedjanova, D., Ferron, F., Beretvas, S. N., & den Noortgate, V. (2020). Effect size estimation for combined single-case experimental designs. *Evidence-based Communication Assessment and Intervention*(14).

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

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

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357-367.

Pustejovsky, J. E. (2018). ARPobservation: Tools for simulating direct behavioral observation recording procedures based on alternating renewal processes [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/ARPobservation/index.html (R package version 1.1)

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217–235. doi: 10.1037/met0000179

Pustejovsky, J. E., & Ferron, J. M. (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. 168-186). New York, NY: Routledge.

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2019). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification.* doi: 10.1177/0145445519864264

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

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

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (pnd) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 163-187. doi: 10.1080/17489530802505412

Scott, R. W., Peters, R. D., Gillespie, W. J., Blanchard, E. B., Edmunson, E. D., & Young, L. D. (1973). The use of shaping and reinforcement in the operant acceleration and deceleration of heart rate. *Behaviour Research and Therapy*, *11*(2), 179-185.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The Quantitative Synthesis of Single-Subject Research Methodology and Validation. *Remedial and Special Education*, *8*(2), 24-33. doi: 10.1177/074193258700800206

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

Shadish, W. R., & 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

Swan, D. M., Pustejovsky, J. E., & Beretvas, S. N. (2020). *Simulation code.* OSF. Retrieved from `https://osf.io/f8bmu`

Todman, J., & Dugard, P. (1999). Accessible randomization tests for single-case and small-n experimental designs in aac research. *Augmentative and Alternative Communication*, *15*(1), 69-82. doi: 10.1080/07434619912331278585

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

Wilson, C. C., Robertson, S. J., Herlong, L. H., & Haynes, S. N. (1979). Vicarious effects of time-out in the modification of aggression in the classroom. *Behavior Modification*, *3*(1), 97-111.

Wolery, M. (2013). A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education*, *34*(1), 39-43. doi: 10.1177/0741932512468038

Table 1

Simulation Conditions

Poisson Data Conditions	
Mean level (μ)	5, 15, 25
Autocorrelation (ϕ)	0, 0.2, 0.4
Gamma Point Process Data Conditions	
Mean level (μ)	5, 15, 25
Dispersion (κ)	5/2, 3/2, 2/3, 2/5
Normally-distributed Data Conditions	
Mean level (μ)	5
Variance (σ^2)	0.25, 1, 2.25
Autocorrelation (ϕ)	0, 0.2, 0.4

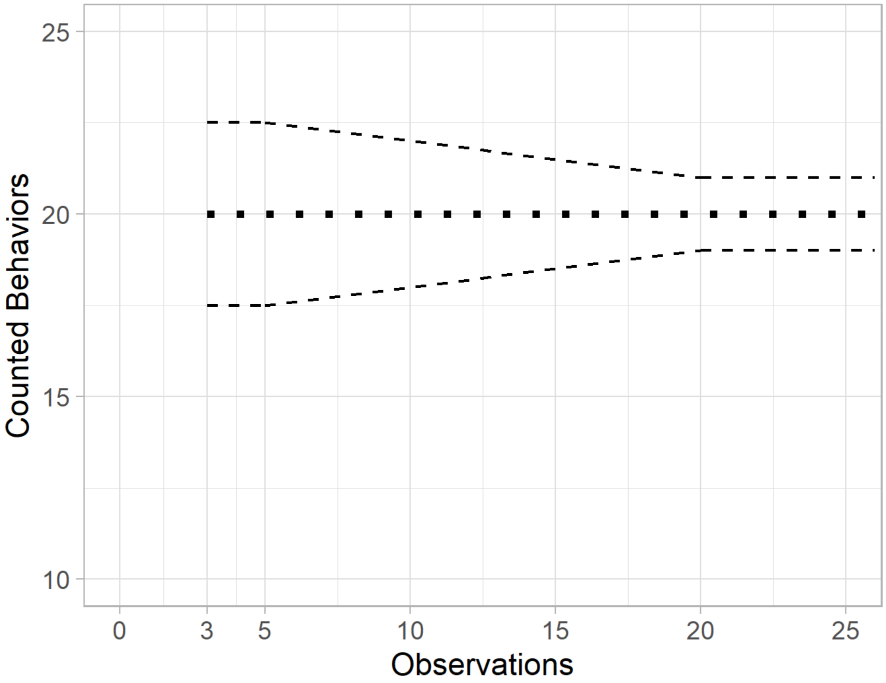

Figure 1. Width of the stability envelope w around a set of observations with a median value of $\tilde{m}_b = 20$. The dotted line represents the value of the median, and the dashed lines represent the bounds of the stability envelope when there are n observations along the x-axis.

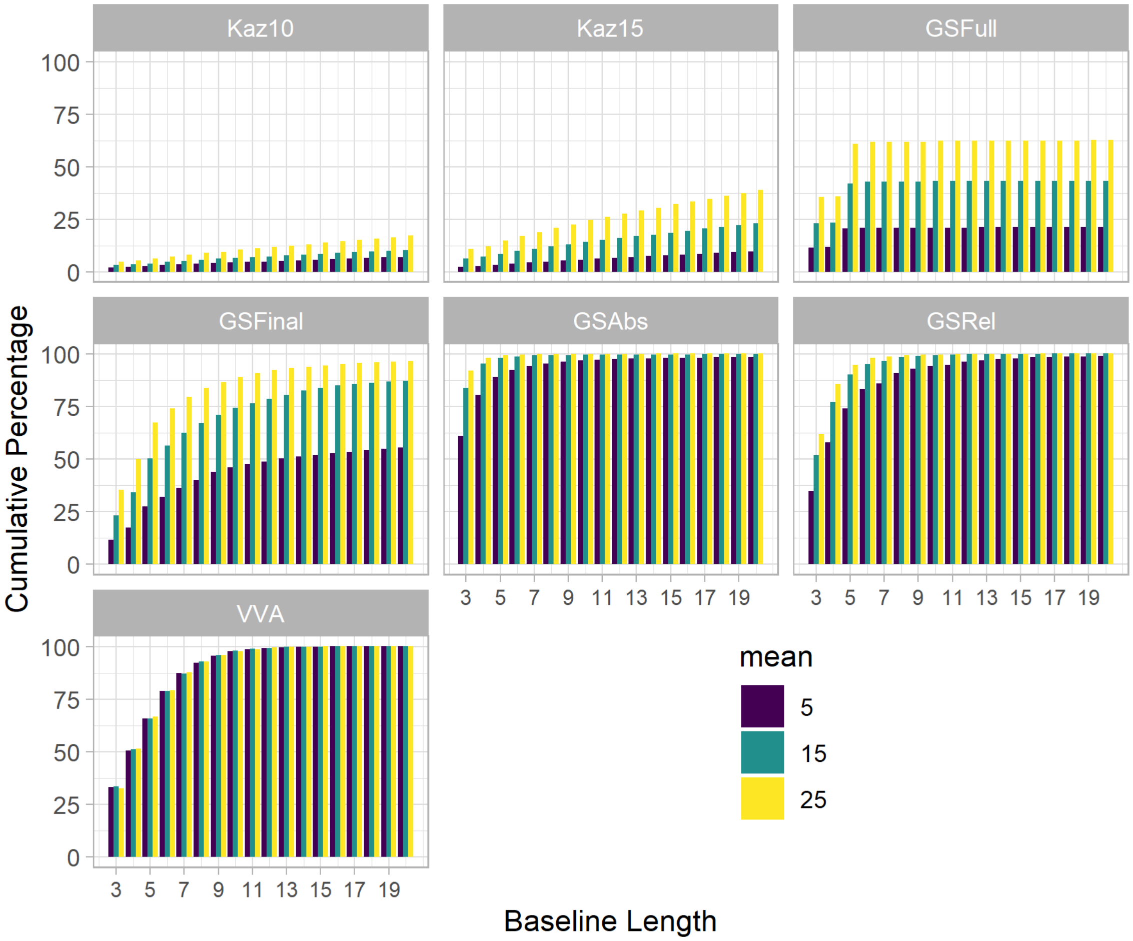

Figure 2. Cumulative percentage of stable baselines at a given number of observations for the independent Poisson case, up to 20 observations.

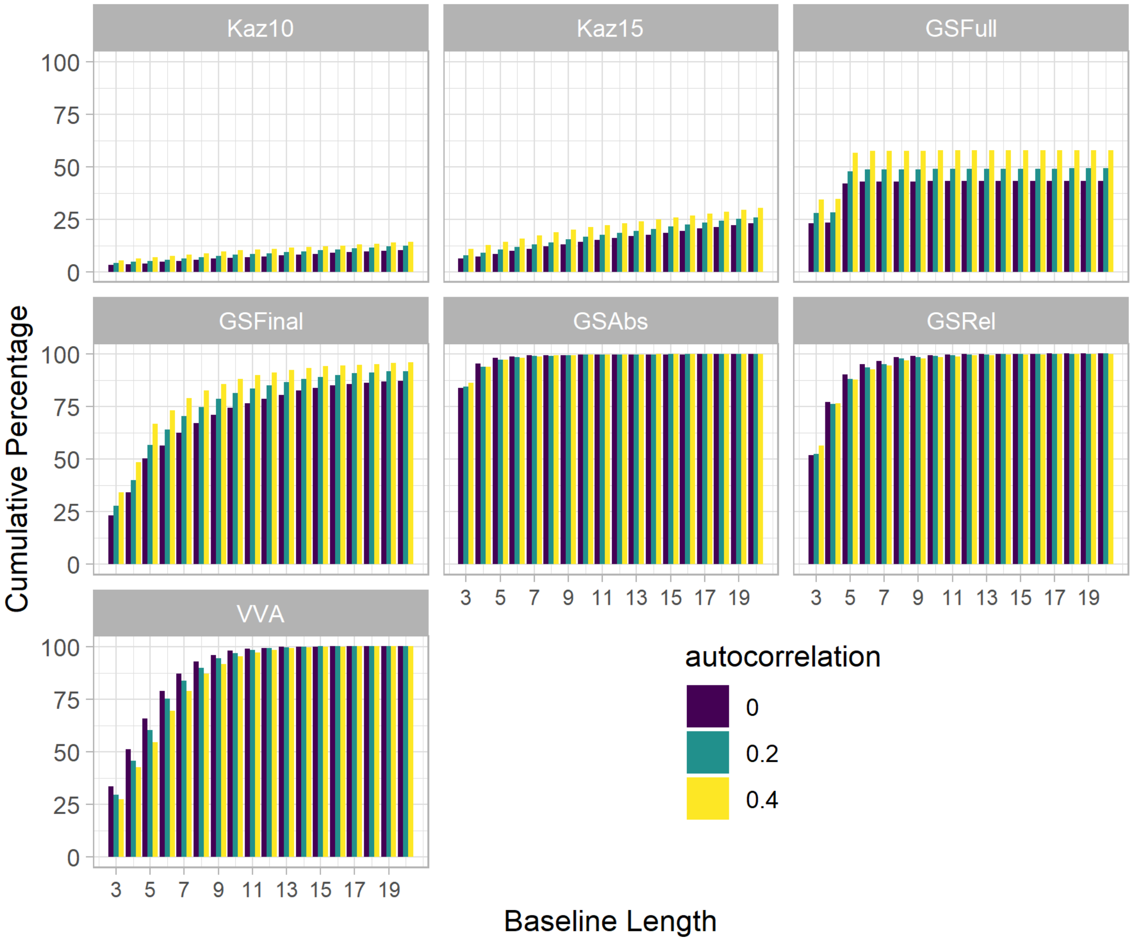

Figure 3. Cumulative percentage of stable baselines at a given number of observations for the Poisson case with mean $\mu = 15$ across different levels of autocorrelation, up to 20 observations.

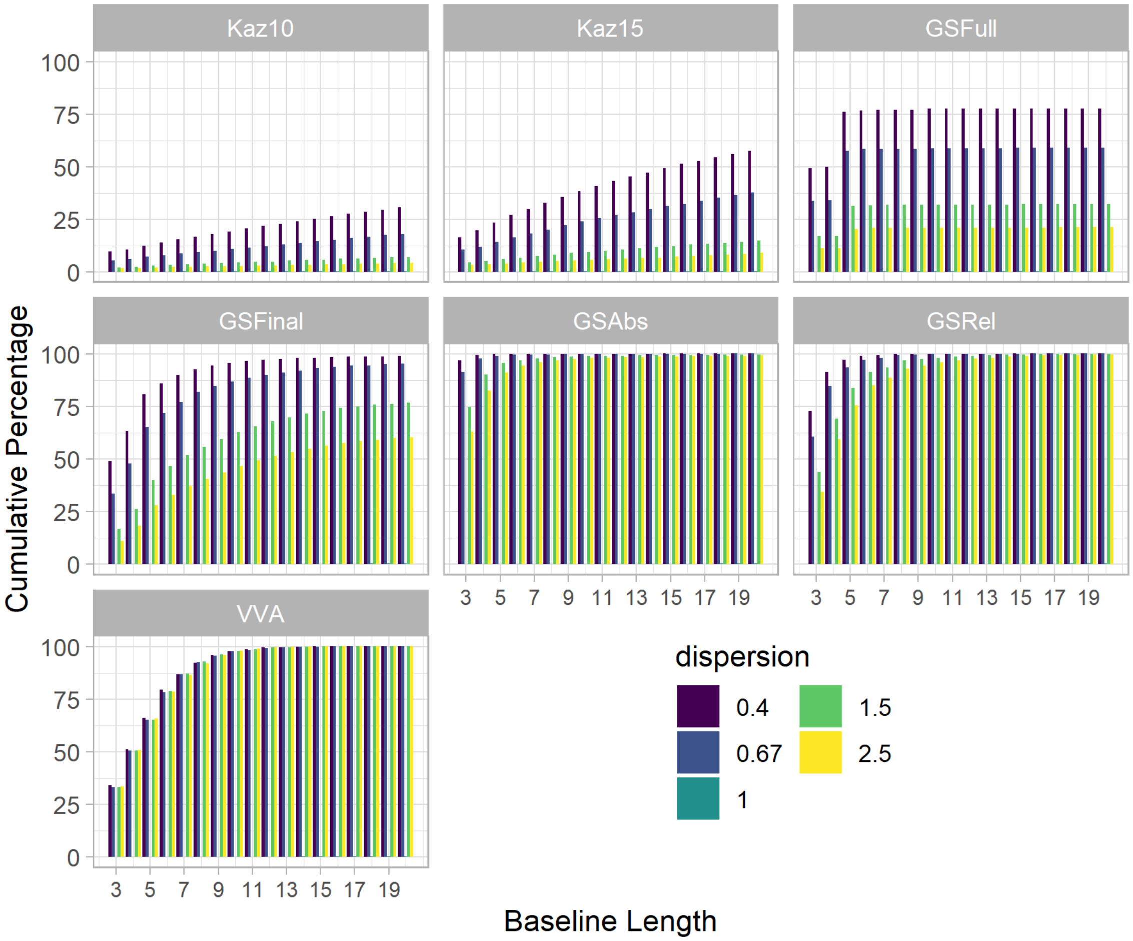

Figure 4. Cumulative percentage of stable baselines at a given number of observations for the gamma point process case with mean $\mu = 15$ across different degrees of dispersion, up to 20 observations.

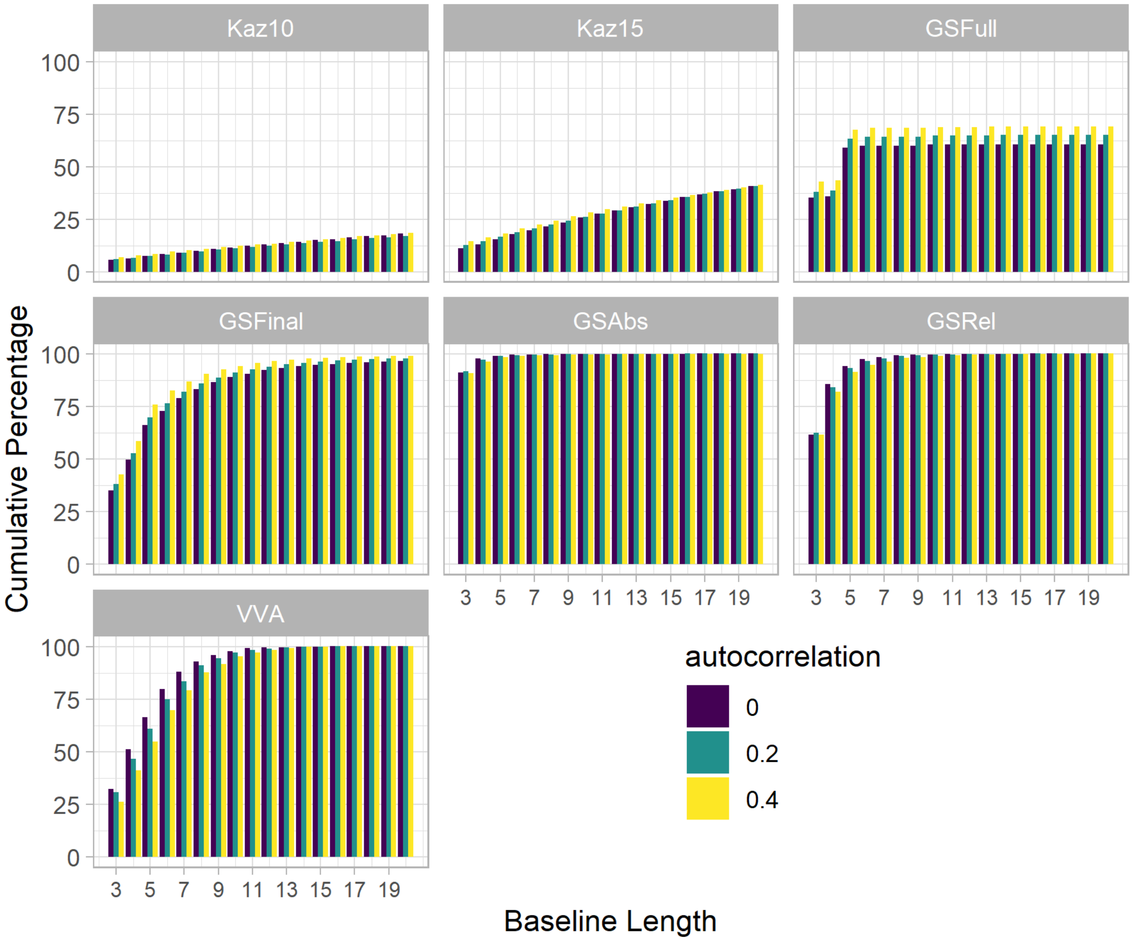

Figure 5. The proportion of normally-distributed replicates designated as a stable baseline across different degrees of autocorrelation when $\sigma^2 = 1$, up to 20 observations.

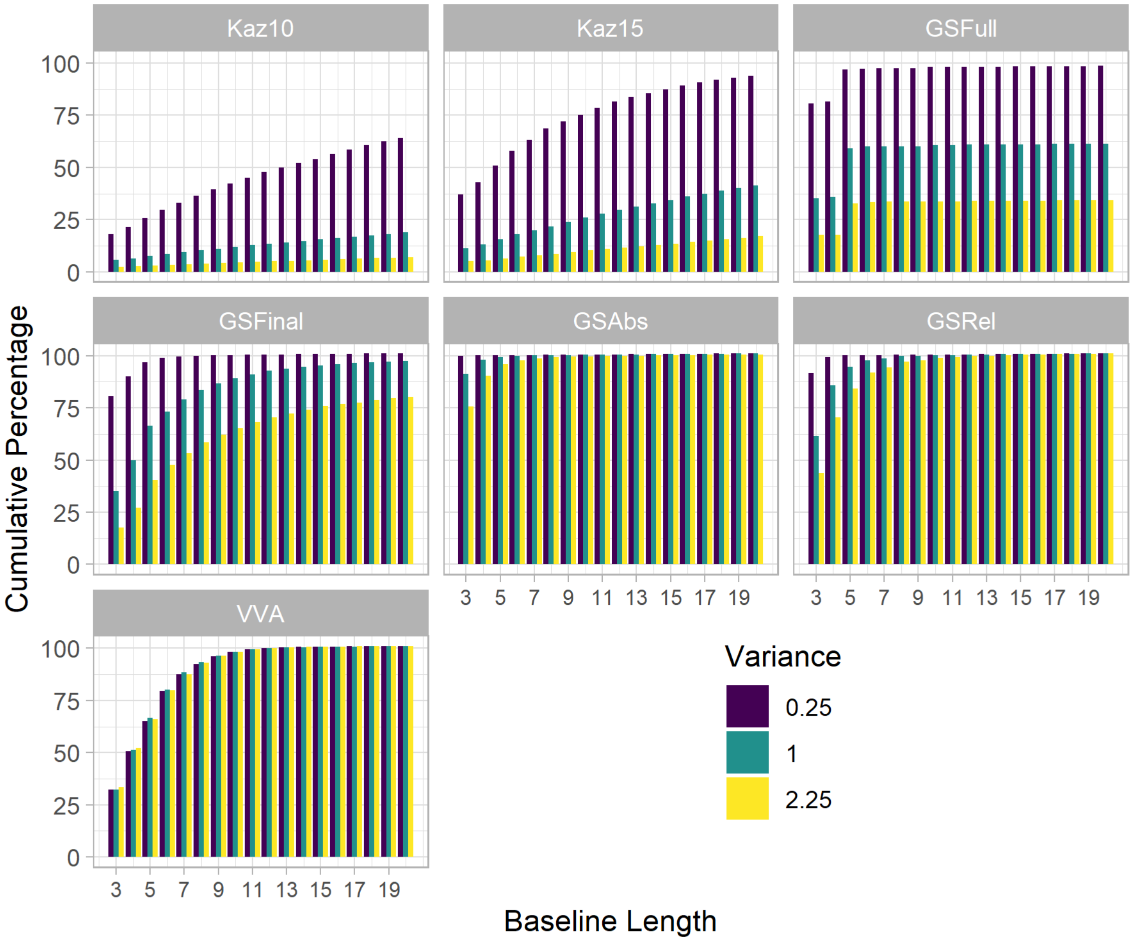

Figure 6. The proportion of normally-distributed replicates designated as a stable baseline across different values of the variance when $\phi = 0$, up to 20 observations.

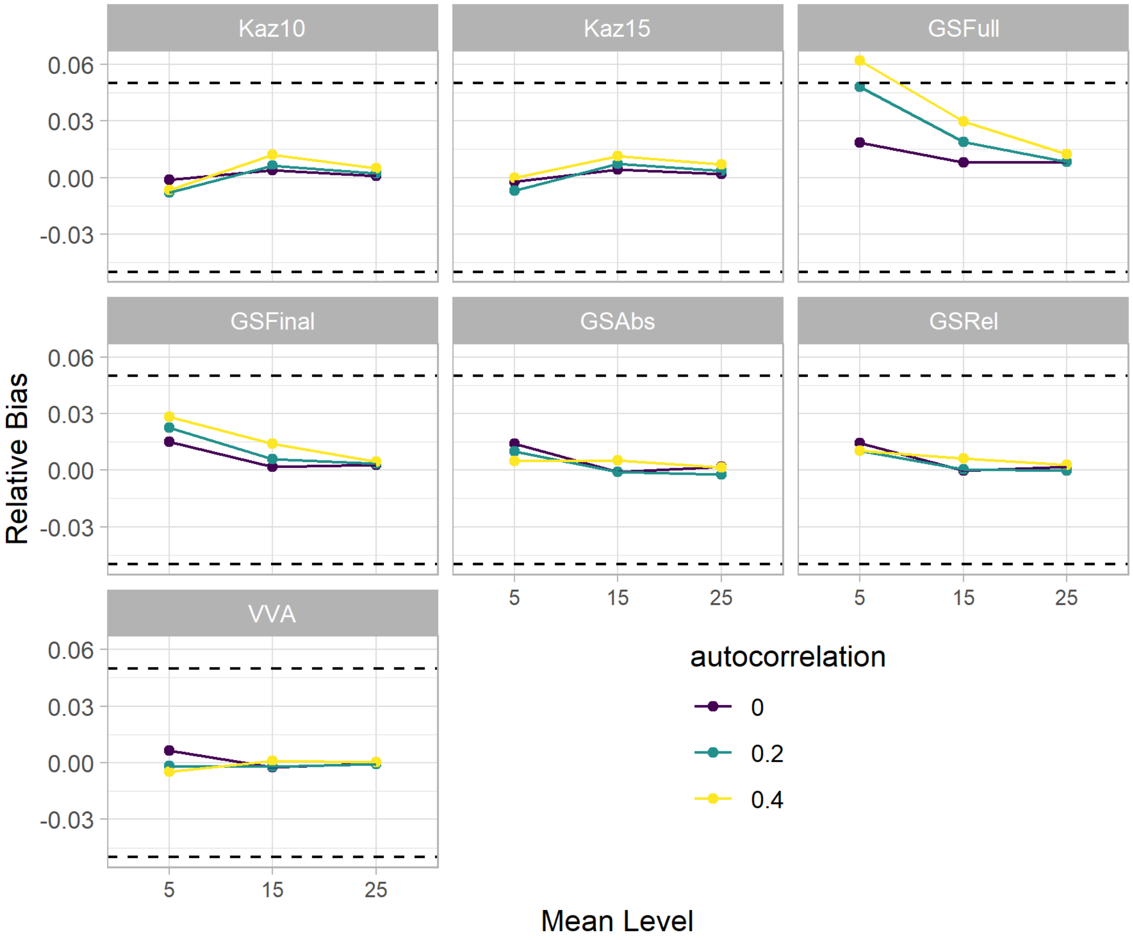

Figure 7. Relative bias of the baseline mean for stable, Poisson-distributed baselines by algorithm and degree of autocorrelation. The dashed lines correspond to relative biases of 5%.

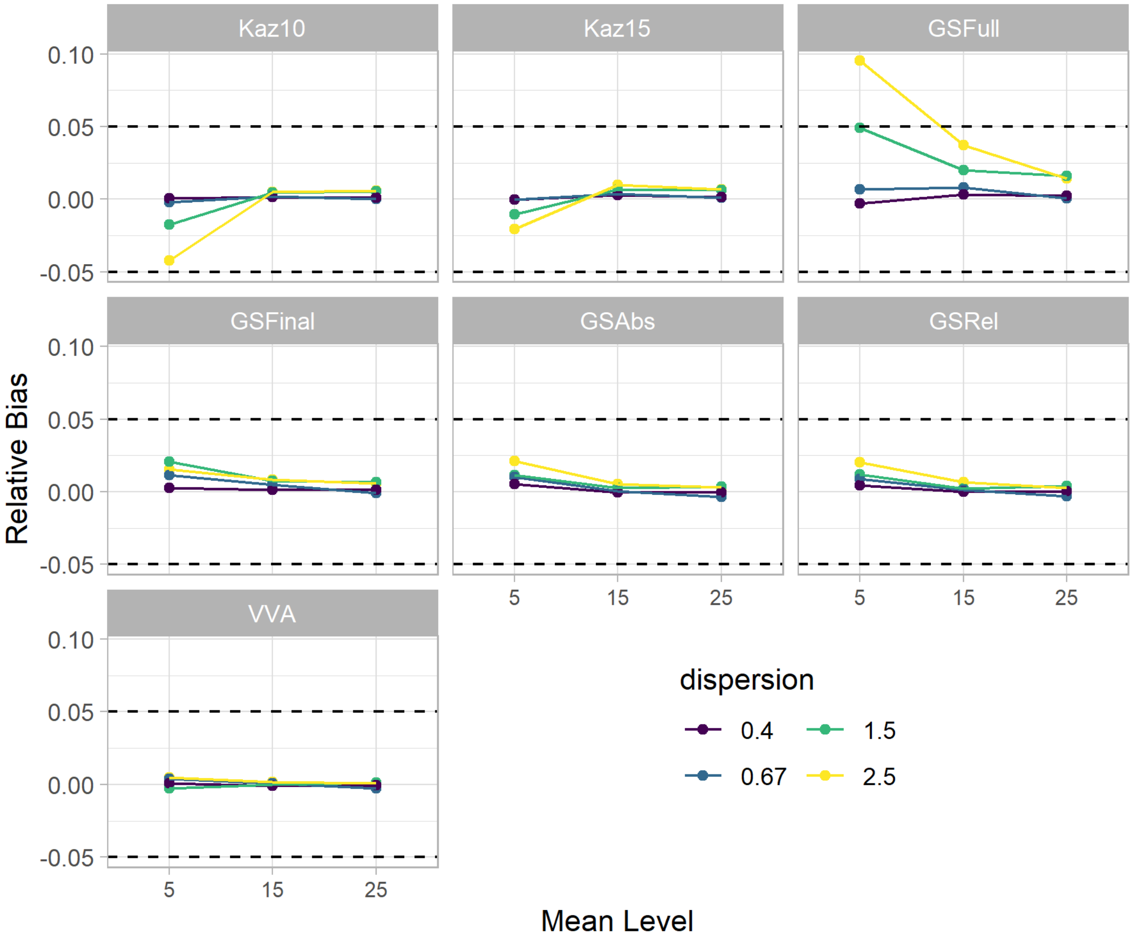

Figure 8. Relative bias of the baseline mean for stable, gamma point process baselines by algorithm and degree of dispersion. The dashed lines correspond to relative biases of 5%.

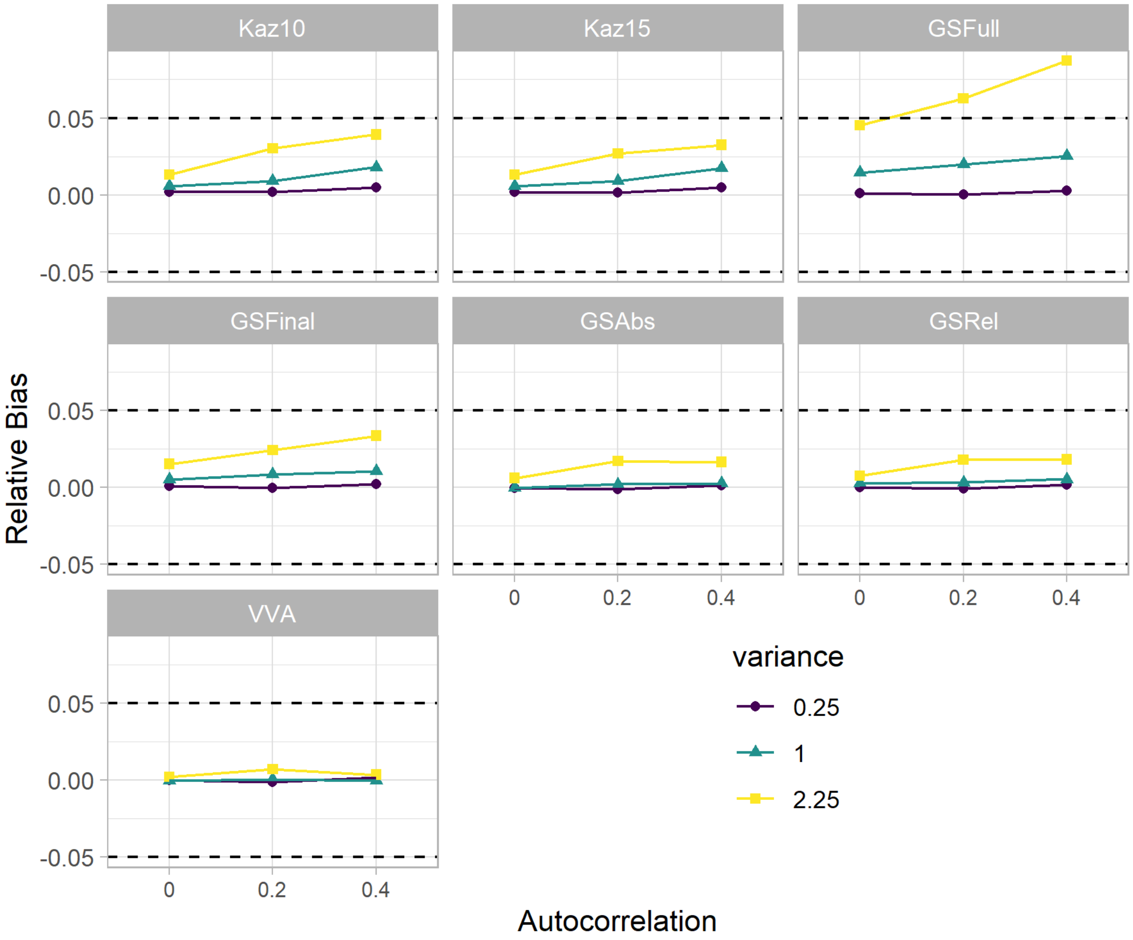

Figure 9. Relative bias of the baseline mean for stable, normally-distributed baselines by algorithm and degree of autocorrelation. The dashed lines correspond to relative biases of 5%.

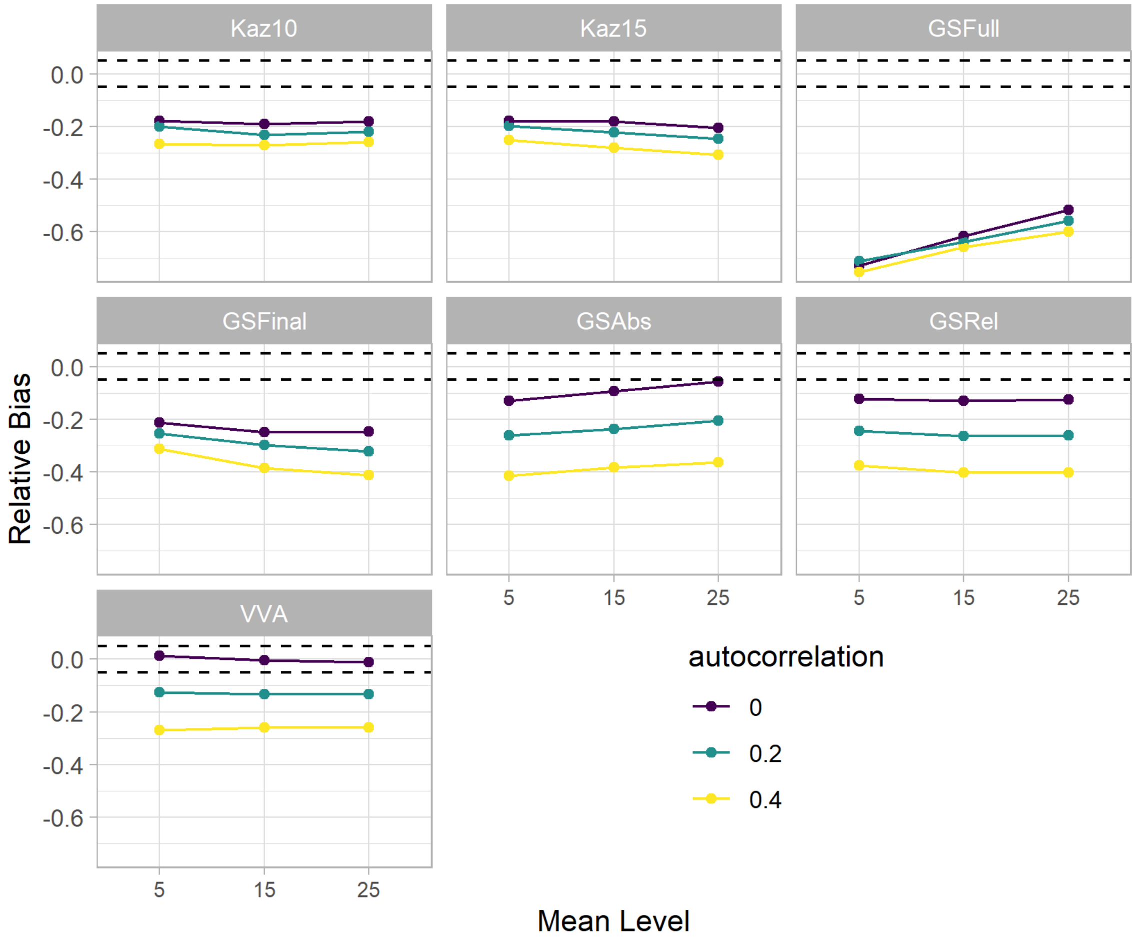

Figure 10. Relative bias of the baseline variance for stable, Poisson-distributed baselines by algorithm and degree of autocorrelation. The dashed lines correspond to relative biases of 5%.

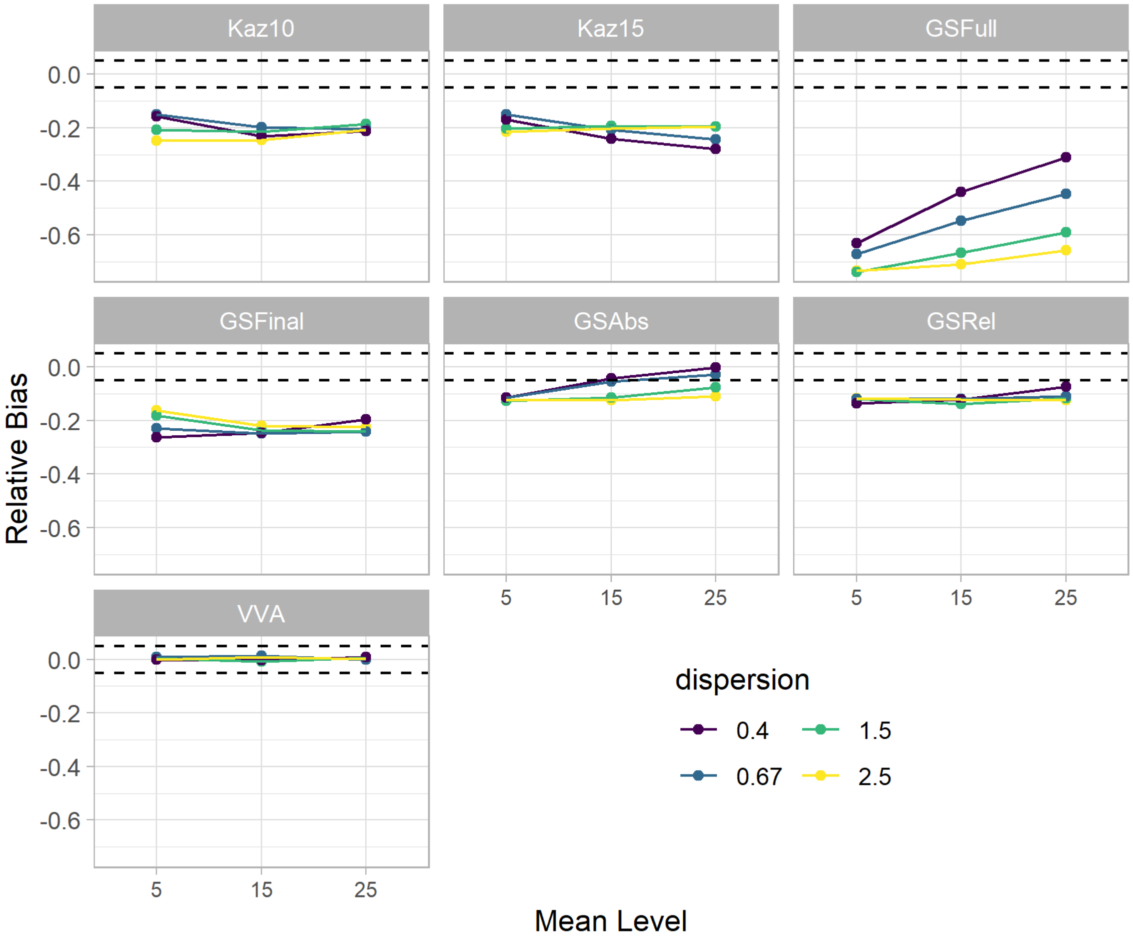

Figure 11. Relative bias of the baseline variance for stable data series following the gamma point process model, by algorithm and degree of dispersion. The dashed lines correspond to relative biases of 5%.

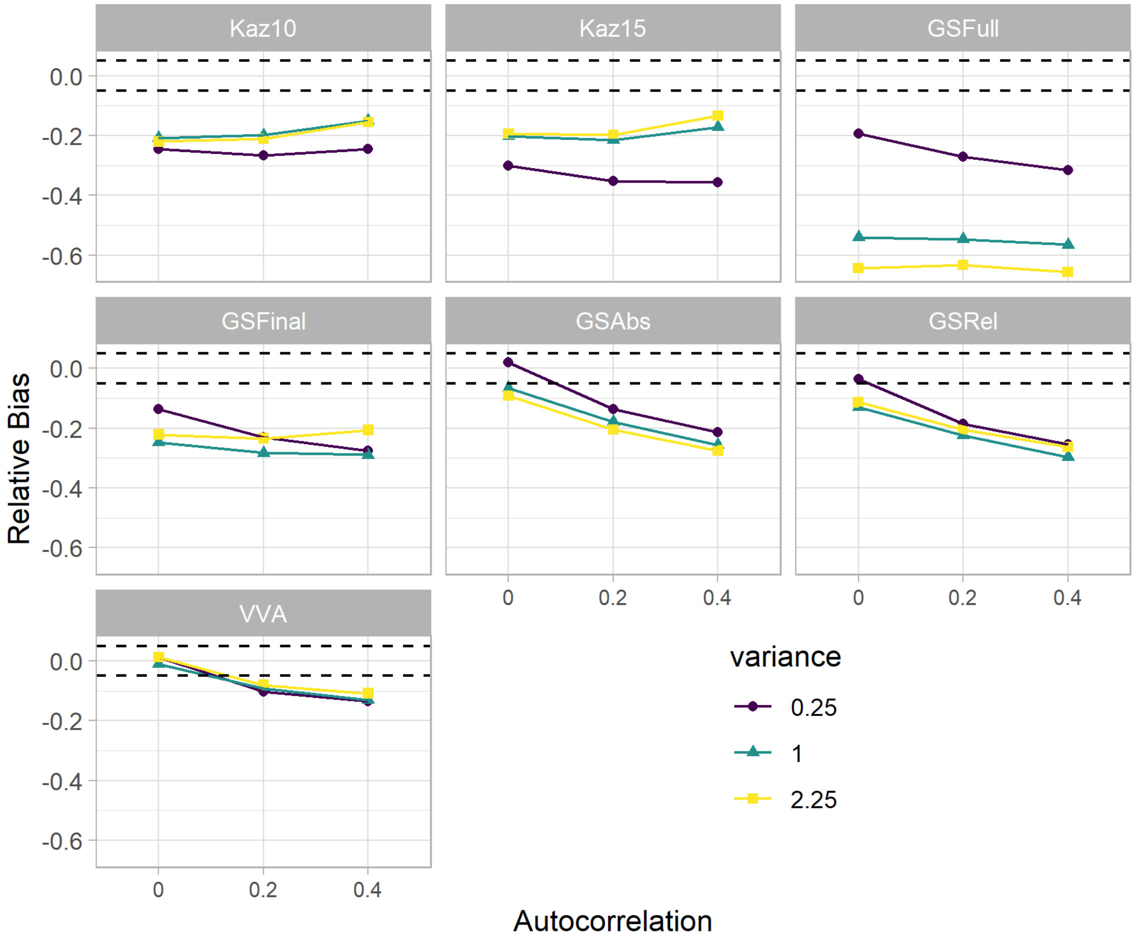

Figure 12. Relative bias of the baseline variance for stable, normally-distributed baselines by algorithm and degree of autocorrelation. The dashed lines correspond to relative biases of 5%.