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

TEXAS Education
The University of Texas at Austin
College of Education

Daniel M. Swan (dswan@utexas.edu), James E. Pustejovsky (pusto@austin.utexas.edu), & S. Natasha Beretvas (tberetvas@austin.utexas.edu)

Overview

There is a history of interest in the meta-analysis of single-case designs (SCDs) going back several decades. Pustjevosky and Ferron (2017) described a number of benefits to the meta-analysis of SCD data, including generalizing the effects of treatments and providing context for treatment heterogeneity. However, for the meta-analytic results to have meaning, we need appropriate models and effect sizes for SCDs that account for SCD-specific researcher practices such as response-guided experimentation.

Response-guided experimentation/designs refer to SCDs where the investigator allows the design of the experiment to be guided by the data on an ongoing basis during a study. Typically, a response-guided design involves making inferences about features of the data. These inferences are not summary inferences about the effectiveness of an intervention, rather they are judgments about patterns in the data that drive decisions about the implementation of an intervention. In this form of response-guided design the analyst is generally concerned with "stability" in the current phase for any or all participants. The specific methods researchers use to establish stability vary, but they generally involve making certain that the data the baseline is a) not too variable
and b) not trending in the expected direction of the effect.

The restriction of variability in the outcome data is an important consideration when applying parametric methods to SCDs. If the variability of the sample data is restricted with respect to the population processes that the sample is drawn from, then it has implications for the estimation of standard errors from statistical models, as well as any effect size that uses the sample standard deviation to standardize the effect.

While the general practice of using response-guided designs is common, the precise set of criteria any given researcher is using in their study is rarely articulated. In this study, we draw on two applied texts (an early influential textbook and a popular, modern textbook) as well as a methodological study for different criteria to characterize stability.

Response Guided Criteria

Source	Abbrev.	General Rules
Kazdin, A. E. (1982). Single-case research designs: methods for clinical and applied settings. New York: Oxford University Press.	• **Kaz10** • **Kaz15**	To characterize a baseline as stable, at last three observations in the baseline must be within either ±5% of the full-baseline mean or ± 7.5% of the full-baseline mean.
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.	• **GSFull** • **GSFinal** • **GSAbs** • **GSRel**	In addition to ensuring that the baseline is not notably trending in the direction of the expected effect, stability of the baseline level is characterized in one of four ways: 1) 80% of the observations in the baseline are within a stability envelope proportional to the magnitude of the median of the baseline observations, 2) the last 3 observations are within a stability envelope proportional to the magnitude of the median of the baseline observations, 3) the magnitude of the difference between the mean of the first half the baseline and the second half of the baseline is not too large, or 4) the difference between the first and last observation in the baseline is not too large.
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.	• **VVA**	Characterizing a stable baseline requires four criteria to be met: 1) the OLS regression slope is less than 0.5 times the SD of the baseline, 2) the OLS slope of the final 3 observations is less than 0.5 the SD of the baseline, 3) the final observation is no more than 2 standard deviations away from the mean of the baseline, and 4) the difference between the mean of the last half of the baseline observations and the first half of the baseline observations is no more than 1.5 standard deviations.

Method

We performed a Monte Carlo simulation, producing simulated SCD baselines 100 observations long . For each response guided criteria, we tested for stability at 3 observations. If the baseline was not stable for a given criteria, we extended the criteria to include the next observation, repeating until we determined how many observations were required for a particular simulated baseline to be considered stable or until we reached the full 100 observations. For each set of conditions (table below) we generated 5000 replicated baselines. We recorded the mean and variance of the **stable** baselines at the first point of stability, separately for each algorithm.

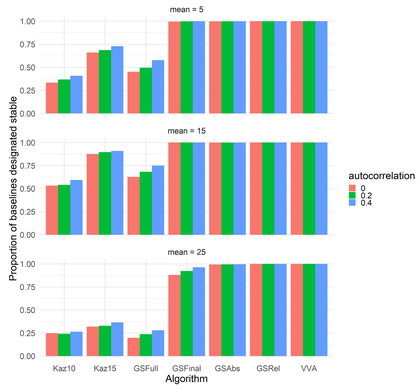
Although the use of non-normal data-generating models is becoming more common in SCD simulations, the use of the normal distribution is still fairly common. In some cases, the implied assumption about the distributions of the outcomes is reasonable. However, in the context of response-guided designs outcome variability is a key consideration in determining stability. Using data generating models that mimic the tightly linked mean-variance relationship of common SCD outcomes, such as counts, is important to understanding the consequences of response-guided designs on data features such as the mean and variance of the outcome. For this simulation, we focused on Poisson-distributed and marginally Poisson-distributed outcomes with an approximate AR(1) structure using a data generating process called binomial thinning (McKenzie, 1988).

Baseline simulation conditions

Mean level (μ)	5, 15, 25
Autocorrelation (ϕ)	0, 0.2, 0.4

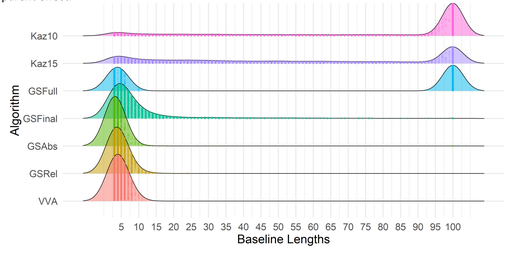
Proportion of replicates characterized as stable

Bar plots of the proportion of replicates found stable, separately by algorithm, generating mean, and generating degree of autocorrelation. When not all baselines reach stability, higher autocorrelation appears to be related to more stable baselines.



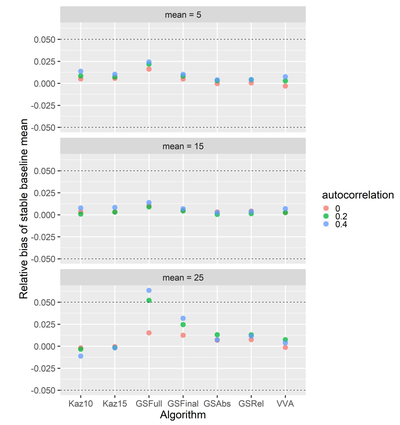
Distribution of baseline lengths

For a generating mean of μ = 15 and ϕ = 0.2. Baselines that were never characterized as stable are truncated at a length of 100. Dots represent individual baselines. Larger generating means lead to shorter baselines. In the case of the Kazdin criteria and the GSFinal criteria, increased autocorrelation appears to yield shorter baselines, if they are stable before 100 observations.. For all other criteria there is no apparent effect.



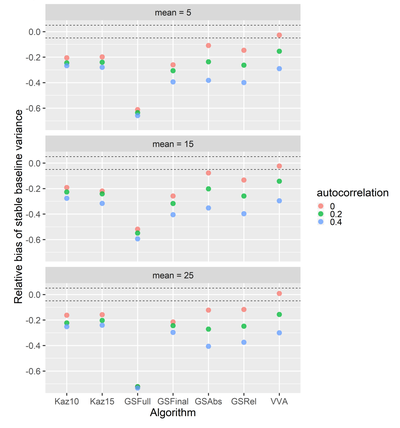
Relative bias of the baseline mean

Plot of the relative bias of the baseline mean, separately by generating mean and generating degree of autocorrelation for only **stable** baselines. The dotted horizontal lines represent ±5% bias with respect to the expected value. Generally speaking, the mean level is unbiased.



Relative bias of the baseline variance

Plots of the relative bias of the baseline variance, separately by generating mean and generating degree of autocorrelation for only stable baselines. The dotted horizontal lines represent ±5% bias with respect to the expected value. For the cases with autocorrelation, we anticipated that the variances will be underestimated. However, for the case where all observations are independent, only the VVA criteria produce stable baselines without bias in the variance. In all other cases the variances are underestimated, especially in the case of the GSFull criteria.



Implications

There are several important implications. The first is that we need to encourage SCD researchers to provide more information about exactly how they are performing SCD-specific practices such as response-guided experimentation. There is likely considerable variability in how different researchers make judgements about stability, and until we can understand the complete range of practices it is difficult to know the true extent of potential biases in existing SCD data.

The second implication is that any statistical model that is applied to data from a response-guided design is likely to have underestimated standard errors, and effect sizes that are standardized by the standard deviation of the outcome are likely to be inflated in magnitude. Researchers interested in models for the analysis and meta-analysis of SCDs need to consider these potential biases as they develop methods for applied researchers.