A hierarchical ordinal regression model for treatment-reversal designs with application to non-overlap effect sizes

James Ohisei Uanhoro and Megan Rojo

Department of Educational Psychology, University of North Texas

Author Note

This work was not supported by any grants. The authors declare no competing interests. All code for simulation studies, data analyses and Stan scripts are available at `https://osf.io/e5cwg/`.

Abstract

We present a hierarchical ordinal model for analyzing single-case designs (SCDs), with a focus on treatment-reversal designs. SCDs involve systematic measurement of outcomes for individual cases across different conditions or phases, aiming to establish causal relations between interventions and behavioral changes. While visual analysis is a common approach in SCDs, the field is increasingly adopting quantitative effect size metrics, such as non-overlap indices, to supplement visual examination. However, statistical theory supporting the use of these indices remains underdeveloped. To address this gap, we developed a Bayesian hierarchical ordinal model that enables the estimation of case-specific non-overlap indices. Through simulation studies, we demonstrate that these indices are more accurate than those obtained via standard approaches. Moreover, the model can generate parametric indices with greater accuracy than standard methods. To facilitate the adoption of this model, we provide an R package (*ssrhom*) for model estimation. This contribution aims to enhance the analysis and interpretation of SCDs, ultimately advancing our understanding of the efficacy of interventions and promoting evidence-based decision-making.

Keywords: single-case design, hierarchical modeling, ordinal regression, effect sizes, non-overlap of all pairs, proportion exceeding median

A hierarchical ordinal regression model for treatment-reversal designs with application to non-overlap effect sizes

Single-case designs (SCDs) represent a crucial methodology in behavioral sciences, particularly in special education and applied behavior analysis. These designs involve systematic measurement of outcomes for individual cases across different conditions or phases, enabling researchers to establish causal relations between interventions and behavioral changes (Horner et al., 2005). The intensive longitudinal nature of SCDs makes them particularly valuable for studying low-incidence populations and evaluating individualized interventions where group designs may be impractical or inappropriate (Gast & Ledford, 2014). Their ability to accommodate individual variability while maintaining experimental control has established SCDs as a cornerstone methodology for developing evidence-based practices in educational and behavioral interventions (Maggin, Lane, & Pustejovsky, 2017).

Visual Analysis and Effect Sizes in Single-Case Research

The analysis of SCD data has historically relied on visual analysis, involving systematic examination of patterns in level, trend, and variability across phases (Gast & Ledford, 2014). While visual analysis remains fundamental to single-case research, the field has increasingly recognized the need for indices or quantitative effect size measures to complement these visual methods (Maggin & Odom, 2014). This shift has been driven by several factors: the need for objective metrics to support meta-analysis, the growing emphasis on effect size reporting in evidence-based practice standards, and the desire for more precise quantification of intervention effects (Maggin, Cook, & Cook, 2019). Among various effect size measures, non-overlap indices have gained particular prominence due to their conceptual alignment with visual analysis principles and their interpretability for practitioners (Parker, Vannest, & Davis, 2011). Non-overlap indices quantify the extent to which data points in one phase (e.g., baseline) do not overlap with data points in another phase (e.g., intervention).

Limitations in Effect Size Inference

Current approaches to computing and interpreting effect sizes in SCDs, particularly non-overlap indices, face several methodological limitations. Measures like percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), non-overlap of all pairs (NAP; Parker & Vannest, 2009), percentage of data points exceeding the median (PEM; Ma, 2006), and Tau-U (Parker, Vannest, Davis, & Sauber, 2011) are widely used, but their utility is constrained by statistical and methodological issues that undermine their accuracy and interpretability. Some of the critical limitations include the inability of most non-overlap indices to produce reliable measures of parameter uncertainty, such as standard errors (SE) or confidence intervals, which are essential for making rigorous inferences about intervention effectiveness and the sensitivity of the effect sizes measures to trends, ceiling effects, or outlying data points (Shadish, 2014a; Shadish, Rindskopf, & Hedges, 2008).

For instance, PND, which measures the percentage of data points in the treatment phase that exceed the highest data point in the baseline phase (Parker, Vannest, & Davis, 2011), is simple to calculate and widely used due to its alignment with visual analysis. However, it focuses on the highest baseline data point, making it highly sensitive to outliers and sensitive to the number of baseline observations (Allison & Gorman, 1994). Moreover, PND lacks statistical rigor as it does not account for trend within phases and does not have a method to calculate SEs or confidence intervals, which limits its application in meta-analytic contexts (Kratochwill et al., 2013; Parker, Vannest, & Davis, 2011).

PEM offers another approach by examining the proportion of intervention data exceeding the median of baseline. While computationally straightforward, PEM similarly suffers from sensitivity to data distributions and lacks methods for estimating SEs, reducing its reliability in studies with skewed or highly variable data (Parker, Vannest, & Davis, 2011).

NAP compares every data point in one phase with every data point in another to assess overlap, with a higher NAP indicating a more effective intervention (Parker,

Vannest, & Davis, 2011). Among non-overlap indices, NAP stands out for its ability to generate confidence intervals through established methods; however, these calculations rely on the assumption of independent observations—a condition that is often violated in SCDs due to autocorrelation or serial dependency in repeated measurements (Barnard-Brak, Watkins, & Richman, 2021). Ignoring this dependency can lead to overly optimistic confidence intervals and biased effect size estimates (Kratochwill et al., 2013).

We intend to address the aforementioned issues in this paper by proposing a model that returns accurate non-overlap effect sizes. However, we focus only on the PEM and NAP (and Tau-U by extension since Tau-U $= 2 \times$ NAP -1). We focus on this subset of non-overlap effect sizes because most other non-overlap effect sizes do not have stable parameter definitions, i.e., these effect sizes, to a large extent, depend on the number of observations in each phase (Pustejovsky, 2019).

We also note that the challenges with estimating non-overlap effect sizes should not cause methodologists to discard them as tools for understanding results in SCD studies. As earlier mentioned, these effect sizes reflect visual analysis, which is fundamental to SCD research. Additionally, there are reasons to believe that non-overlap effect sizes are easier to understand than standard parametric effect sizes. For example, McGraw and Wong (1992) describe the NAP as a *common language effect size* as it is 'better than the available alternatives for communicating effect size to audiences untutored in statistics.' In fact, when Cohen presented the standardized mean difference (SMD), he converted the SMD to percentage overlap to make the SMD more 'intuitively compelling and meaningful' (Cohen, 1988, sec. 2.2.1). Finally, there is also empirical evidence to suggest that non-overlap effect sizes are easier to understand than standard parametric effect sizes (Brooks, Dalal, & Nolan, 2014).

Hierarchical Ordinal Modeling for Computing ES

Although researchers have proposed various solutions to challenges with computing effect sizes for SCD (e.g., Rindskopf, 2014b; Shadish, Zuur, & Sullivan, 2014;

Swaminathan, Rogers, & Horner, 2014), there remains a need for a method to address these limitations simultaneously (Shadish, 2014a). An ideal method would (a) account for data patterns, including time trends, (b) incorporate all available data, (c) rely on appropriate distributional assumptions, and (d) address serial dependence in repeated measurements (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Shadish, 2014b; Swan & Pustejovsky, 2018; Wolery, Busick, Reichow, & Barton, 2010). In response to this need, we proposed and tested a Bayesian hierarchical ordinal model for computing SCD effect sizes. Although our model addresses many of these challenges, it does not address time trends. However, as this is an initial presentation of this approach, we follow precedent in omitting trend effects at this stage, as seen in other foundational work (e.g. Hedges, Pustejovsky, & Shadish, 2013; Pustejovsky, 2018).

Below, we discuss aspects of the model that lead us to believe our approach is promising for analyzing SCDs.

Bayesian modeling. First, we apply Bayesian modeling to *augment the information in likelihood* or data (Levy & McNeish, 2023). SCD data tend to be limited in sample size, resulting in relatively weak information in the data. For this reason, modelling such data can lead to inefficient estimates. To elaborate, a model for such data may result in unbiased estimates, but the data are limited, so the estimates are noisy or imprecise. Hence, we use priors to stabilize the estimation of model parameters. These priors might bias parameters, but overall, the increased stability of parameters should lead to more precise estimates in each sample and more accurate estimates on average.

Hierarchical nature of model. Second, the hierarchical nature of the model enables more precise estimation of effect sizes through principled information sharing across cases, while providing case-specific estimates. This is a common feature of methodological recommendations for analyzing SCD (e.g., Hedges et al., 2013; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Rindskopf, 2014a; Shadish, 2014b; Shadish et al., 2014).

Ordinal regression. Third, our approach leverages a flexible ordinal framework, which respects the ordinal nature of most behavioral measurements (Cliff, 1993), while accommodating autocorrelation structures common in SCD data. SCD data are often limited, making it difficult to verify distributional assumptions. For example, SCD data may be bounded counts. An example is the study by Tasky, Rudrud, Schulze, and Rapp (2008), which counted the number of times, across six trials, that individuals with traumatic brain injury remained on task. Therefore, the data were guaranteed to be counts between 0 and 6 inclusive. There are several candidate models for such data: binomial, beta-binomial, or binomial with observation-level random effects. In such situations, the choice of model impacts the results, as data generated according to one model can yield inaccurate inferences when analyzed by another model (Harrison, 2015). A similar phenomenon occurs with unbounded counts (e.g., the number of times a participant displayed a behavior within a set time interval), where using the incorrect distribution leads to inaccurate inference (Harrison, 2014; Wooldridge, 2010). In contrast, ordinal regression handles issues that arise with counts, such as under- or over-dispersion, boundedness, and heaping, because it directly models the conditional distribution of the data (Kowal & Wu, 2023; Valle, Ben Toh, Laporta, & Zhao, 2019). In general, for a given dataset, there will be several potential reasonable parametric distributions, and SCD data may be too weak to identify the correct one, assuming the correct choice is among the candidates. Ordinal regression functions as a semi-parametric regression approach that can model the conditional distribution of the data, regardless of its shape (Liu, Shepherd, Li, & Harrell, 2017; Valle et al., 2019). Moreover, the proportional odds model, the most common ordinal model, is equivalent to the standard non-parametric Mann-Whitney test in the two-group comparison case (Whitehead, 1993). If we trust the Mann-Whitney test in the two-group case for atypical data, then ordinal regression provides a way to generalize the robustness of the Mann-Whitney test beyond the two-group situation. The major requirement for the

adequacy of ordinal regression is that the data are unidimensional.[1] If the data are unidimensional, then they are also ordinal, but may not be any of the parametric distributions routinely used in applied statistical analysis. The impact of distributional assumptions especially matters to us because we are using Bayesian modeling, which adheres to the likelihood principle (Berger & Wolpert, 1988): only information transmitted through the likelihood impacts the analysis. Therefore, the wrong choice of distribution would easily yield inadequate Bayesian inference, without clear options for robust-variance corrections (MacKinnon & White, 1985) or quasi-likelihood adjustments (Wedderburn, 1974) common in frequentist analysis. In summary, SCD data are often limited, making it difficult to determine the correct distribution; under such circumstances, our approach to ordinal regression should yield reliable estimates when applied to SCD data.

Effect size estimation. Finally, since we expect the model to accurately capture the distribution of the data, we can confidently compute several effect sizes after model estimation. This includes standard parametric effect sizes such as the SMD, and the non-overlap effect sizes that we focus on, the NAP and PEM. Additionally, Bayesian modeling allows us to obtain the distributions for each effect size, enabling more nuanced inference and uncertainty estimation. The choice of ordinal regression is again important here. It is possible to compute non-overlap measures from standard parametric models (e.g., Kotz, Lumelskii, & Pensky, 2003; McGraw & Wong, 1992), but the measures will be inaccurate if the parametric model is wrong (Li, 2015; Vargha & Delaney, 2000).

Overall, the theoretical validity of this approach rests on two key assumptions: (1) cases within a study share enough similarities to inform each other's estimates while maintaining individual differences, and (2) behavioral measurements, when assumed unidimensional, are also ordinal. While the primary focus of this study was on treatment reversal designs (e.g., ABAB), the framework has potential applications to other single-case

[1] One violation of undimensionality would be data arising from mixtures, e.g., data with zero inflation. With zero-inflation, ordinal regression assumes a hurdle process as opposed to a mixture model.

design structures such as multiple baseline designs.

All code for simulation studies and data analyses is available at `https://osf.io/e5cwg/`.

Hierarchical ordinal model

Data generation process

We begin by assuming a multivariate normal latent variable, z_{ct}, for case $c \in \{1, \ldots, C\}$ at time $t \in \{1, \ldots, T\}$:

$$\begin{aligned} \mathbf{z}_c &\sim \mathcal{N}_T(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}),\ \mu_{ct} = \gamma_c + \beta_c \cdot x_{ct},\ \sigma_{ij} = \rho^{|i-j|} \text{ for } i, j \in \{1, \ldots, T\}, \\ \gamma_c &\sim \mathcal{N}(\gamma_0, \tau_\gamma),\ \beta_c \sim \mathcal{N}(\beta_0, \tau_\beta) \end{aligned} \tag{1}$$

where $x_{\cdot\cdot}$ is an indicator for a treatment phase measurement. The case-specific mean on the latent scale is dependent on a normal hierarchical model, with separate random effects on both the case baseline level, γ_c, and the case treatment effect, β_c. Precisely, we assume a single treatment effect across AB pairs. The data within a case are autocorrelated at ρ, and the scale matrix, $\boldsymbol{\Sigma}$, has unit diagonal for model-identification purposes.

Let y_{ct} represent the observed outcome. We assume that y_{ct} is a discretized version of z_{ct}, and can take on one of R unique values, denoted by $\{u_1, u_2, \ldots, u_R\}$, arranged in increasing order such that $u_1 < u_2 < \cdots < u_R$. We relax the proportional odds assumption of common ordinal models and permit separate thresholds by treatment phase, allowing different distributions for control and treatment data (McCullagh, 1980; Peterson & Harrell, 1990). Hence, the discretization process is based on the value of x_{ct}, which determines which vector of thresholds, $\boldsymbol{\kappa}^{(0)} = \{\kappa_1^{(0)}, \ldots, \kappa_{R-1}^{(0)}\}$ or $\boldsymbol{\kappa}^{(1)} = \{\kappa_1^{(1)}, \ldots, \kappa_{R-1}^{(1)}\}$, is applied. Specifically, if $x_{ct} = 0$, we use $\boldsymbol{\kappa}^{(0)}$, and if $x_{ct} = 1$, we use $\boldsymbol{\kappa}^{(1)}$. The discretization process follows:

$$y_{ct} = \begin{cases} u_1 & \text{if } z_{ct} \leq \kappa_1^{(x_{ct})} \\ u_2 & \text{if } \kappa_1^{(x_{ct})} < z_{ct} \leq \kappa_2^{(x_{ct})} \\ \vdots & \vdots \\ u_{R-1} & \text{if } \kappa_{R-2}^{(x_{ct})} < z_{ct} \leq \kappa_{R-1}^{(x_{ct})} \\ u_R & \text{if } z_{ct} > \kappa_{R-1}^{(x_{ct})} \end{cases} \tag{2}$$

Equations 1 and 2 provide an almost complete data generation process for SCD data. The data for each case are discrete realizations of a multivariate normal distribution. The cases may have different baselines and treatment effects, but have the same structured autocorrelation matrix. The data are then discretized using separate thresholds for the control and treatment phases. This DGP is not complete because the origin of the unique values, $u_.$, is unclear. One can think of them as impacts of the measurement instrument. For example, assuming the unique values represent counts of a behavior, one can describe the propensity for that behavior as latent (z_{ct}), with higher propensities resulting in more instances of the behavior. In practice, we do not generate data according to this model; we use it only to analyze real data for which the unique values are already available.

However, this model should be a fairly robust base for estimating average baseline and treatment levels in multi-case treatment-reversal data, under the simplifying assumption of no time trend within phase. In practice, the data are transformed to (dense) ranks after which the multivariate probit model above can be applied. The model is able to capture differences by treatment phase in location (given β_c) and distributions, thanks to the use of two threshold vectors. The model is also able to capture autocorrelation (given ρ) under the assumption that autocorrelation is identical across phases and cases, though this assumption may be relaxed.

Computing effect sizes

The model can be used to compute several effect sizes of interest to single-subject researchers. To compute these effects, we first need to compute the case-wise probability mass function (PMF) in each phase:

$$\Pr(y_{ct} = u_r \mid x_{ct} = d) = \begin{cases} \Phi\left(\kappa_1^{(d)} - \gamma_c - d \cdot \beta_c\right) & \text{if } r = 1 \\ \Phi\left(\kappa_r^{(d)} - \gamma_c - d \cdot \beta_c\right) - \Phi\left(\kappa_{r-1}^{(d)} - \gamma_c - d \cdot \beta_c\right) & \text{if } 1 < r < R \\ 1 - \Phi\left(\kappa_{R-1}^{(d)} - \gamma_c - d \cdot \beta_c\right) & \text{if } r = R \end{cases} \quad (3)$$

where d is either 0 (control) or 1 (treatment), and $\Phi(\cdot)$ is the standard normal distribution function. Effect sizes depend on the distribution of the data, so researchers can compute any effect size of interest from the case-wise and phase-wise PMFs. We show some examples below.

Means and quantities based on means. The phase mean for case c is given by

$$\mathbb{E}[y_{ct} \mid x_{ct} = d] = \sum_{r=1}^{R} \left[u_r \cdot \Pr(y_{ct} = u_r \mid x_{ct} = d)\right]$$

The case-wise mean difference may then be calculated. The phase variance for case c can also be computed as:

$$\text{Var}(y_{ct} \mid x_{ct} = d) = \sum_{r=1}^{R} \left[(u_r - \mathbb{E}[y_{ct} \mid x_{ct} = d])^2 \cdot \Pr(y_{ct} = u_r \mid x_{ct} = d)\right]$$

The phase means and variances can be combined to compute case-wise standardized mean differences. When $u_1 \geq 0$ such as for counts, the log rate ratio – assuming an increase in the rate is desirable – is (e.g., Pustejovsky, 2015):

$$\text{Log rate ratio} = \log\left(\mathbb{E}[y_{ct} \mid x_{ct} = 1]\right) - \log\left(\mathbb{E}[y_{ct} \mid x_{ct} = 0]\right)$$

When $u_1 \geq 0$ and $u_R \leq 1$ such as for proportions, the log odds ratio is (e.g.,

Pustejovsky, 2015):

$$\text{Log odds ratio} = \text{Log rate ratio} + \log\left(1 - \mathbb{E}[y_{ct} \mid x_{ct} = 0]\right) - \log\left(1 - \mathbb{E}[y_{ct} \mid x_{ct} = 1]\right)$$

Medians. The phase median for case c is:

$$\text{Median}(y_{ct} \mid x_{ct} = d) = \begin{cases} u_1 & \text{if } \Pr(y_{ct} = u_1 \mid x_{ct} = d) > 0.5 \\ u_r & \text{where } \Pr(y_{ct} \leq u_r \mid x_{ct} = d) \geq 0.5 \text{ and } \Pr(y_{ct} \leq u_{r-1} \mid x_{ct} = d) < 0.5 \end{cases}$$

after which quantities like the case-wise log ratio of medians or median differences can be calculated.

Non-overlap of all pairs. Assuming an increase is desirable, the case-wise non-overlap of all pairs (NAP) would be (e.g., Vargha & Delaney, 2000):

$$\text{NAP} = \sum_{r=1}^{R} \left[\begin{aligned} &\Pr(y_{ct} = u_r \mid x_{ct} = 0) \cdot \Pr(y_{ct} > u_r \mid x_{ct} = 1) \\ &\quad + 0.5 \cdot \Pr(y_{ct} = u_r \mid x_{ct} = 0) \cdot \Pr(y_{ct} = u_r \mid x_{ct} = 1) \end{aligned} \right]$$

Proportion exceeding median. Assuming an increase is desirable, the case-wise proportion exceeding the median (PEM) would be:

$$\begin{aligned} \text{PEM} = &\Pr(y_{ct} > \text{Median}(y_{ct} \mid x_{ct} = 0) \mid x_{ct} = 1) \\ &+ 0.5 \cdot \Pr(y_{ct} = \text{Median}(y_{ct} \mid x_{ct} = 0) \mid x_{ct} = 1) \end{aligned}$$

Model estimation

Handling large R

When the number of unique values in the data (R) is large, estimating the threshold vectors, each consisting of $R - 1$ elements, becomes challenging. To address this issue, we draw on techniques from the literature on ordinal models for continuous data (Manuguerra & Heller, 2010; Manuguerra, Heller, & Ma, 2020). Specifically, we use I-splines (Ramsay, 1988) to approximate the thresholds, which helps enforce the monotonicity constraint on

the threshold vectors. The spline basis functions will have as many rows as there are elements in the threshold vector. The number of columns, or degrees of freedom (p), indicates the complexity of the splines, with more columns allowing for the representation of more complex functions. Additionally, the degree of the splines (n) captures their smoothness: a spline of degree zero is a step function, while a spline of degree two is piecewise quadratic. To estimate the threshold vector, we assume:[2]

$$\begin{cases} n = 0,\ p = 0 & \text{if } R = 2 \\ n = 0,\ p = R - 1 & \text{if } 3 \leq R \leq 10 \\ n = 2,\ p = R - 3 & \text{if } 11 \leq R \leq 20 \\ n = 2,\ p = 18 & \text{if } R > 20 \end{cases}$$

For binary outcome variables, the lone threshold is assumed to be 0, and we do not estimate the threshold vectors. For datasets with 10 or fewer unique data points, we estimate all thresholds directly. For datasets with more unique data points, we use quadratic splines to approximate the thresholds. We assume that 18 degrees of freedom are always sufficient to approximate the threshold vector when there are more than 20 unique data points.

Bayesian estimation

Given the complexity of the model at hand, we employ Bayesian estimation to *augment the information in the likelihood* (Levy & McNeish, 2023) such that parameter

[2] Since we relax the proportional odds constraint, we effectively have two intercepts: γ_c for the control phase, and $\gamma_c + \beta_c$ for the treatment phase. In ordinal regression, estimating the regression intercept requires estimating one less threshold parameter. This is because all thresholds and the intercept cannot be jointly identified unless there is a constraint on the thresholds. However, since we estimate the thresholds using I-splines, and the left boundary of I-splines is guaranteed to be 0, an intercept is necessary to adjust the level of the I-splines. We centered the I-splines basis functions at 0.5 to ease estimation efficiency. This centering means the left boundary of the curve will not necessarily be zero, potentially eliminating the need for an intercept. However, since we penalized the spline weights (see the section on Bayesian estimation on page 11), we need an intercept to ensure the curve is flexible enough to capture extremely high initial or extremely low final thresholds. Such extreme initial or final thresholds can arise when most of the observed responses are either the smallest or the largest unique response value.

estimates are more efficient than they would be in the absence of any prior information. We now describe the prior choices for parameters:

$$\begin{gathered}\gamma_c \sim \mathcal{N}(\gamma_0, \sigma_\gamma),\ \beta_c \sim \mathcal{N}(\beta_0, \sigma_\beta),\ \gamma_0 \sim \mathcal{N}(0, 5),\ \beta_0 \sim \mathcal{N}(0, 2.5),\ [\sigma_\gamma, \sigma_\beta] \sim t^+(3), \\ \frac{\rho + 1}{2} \sim \text{Beta}(2, 2),\ \boldsymbol{\xi}^{(0)} \sim \mathcal{N}^+(0, \sigma_\xi),\ \boldsymbol{\xi}^{(1)} \sim \mathcal{N}^+(0, \sigma_\xi),\ \sigma_\xi \sim \mathcal{N}^+(0, 1)\end{gathered} \tag{4}$$

where $\boldsymbol{\xi}^{(0)}$ and $\boldsymbol{\xi}^{(1)}$ are p-dimensional coefficients vectors for the control and treatment spline functions, respectively. Consequently, $\kappa_r^{(0)} = \mathbf{s}_r' \cdot \boldsymbol{\xi}^{(0)}$ and $\kappa_r^{(1)} = \mathbf{s}_r' \cdot \boldsymbol{\xi}^{(1)}$, where $\mathbf{s}_r'$ denotes the r-th row of the spline basis functions.

The priors in equation 4 are mostly weakly informative priors (Lemoine, 2019) that aim to keep the parameters within realistic ranges given the probit response scale. The priors on $\boldsymbol{\xi}^{(\cdot)}$, however, are intended to regularize the spline weights towards zero, improving their accuracy (Ruppert, Wand, & Carroll, 2003, section 5.5).

We believe these priors will be adequate for most scenarios, but we also encourage researchers to make them more informative when substantial prior evidence is available. Examples include more informative priors on the autocorrelation coefficient, ρ, or the average treatment effect, β_0.[3] β_0 is akin to a standardized mean difference with the residual standard deviation as the effect size denominator, since the model assumes that the regression residual standard deviation is one ($\boldsymbol{\Sigma}$ has a unit diagonal).

One advantage of using a Bayesian approach is that effect sizes can be computed within each iteration of posterior sampling, thus we can realize the distribution of effect sizes. The model above and effect sizes are implemented in the *ssrhom* package in R (Uanhoro, 2025). We now turn to data demonstration exercises to illustrate the model's behavior with SCD data.

---

[3] Although it is possible to modify the Bayesian code to use more informative priors, the package does not currently support this feature, but it will in future versions.

Data demonstrations

We present two data demonstrations with ABAB designs, using data from Tasky et al. (2008) and Shogren, Faggella-Luby, Bae, and Wehmeyer (2004). Tasky et al. (2008) data are the number of times out of six that participants were able to stay on-task and are discrete. The Shogren et al. (2004) study is a systematic review of nine studies on the effect of choice-making, compiled and reanalyzed by Pustejovsky (2015) and available in the *SingleCaseES* R package (Pustejovsky, Chen, & Swan, 2023); most of the outcome measures were proportions.

We analyze the data using the hierarchical ordinal model (HOM) above. We examine effect size convergence using the $\widehat{R}$ statistic (preferably below 1.01) and effective sample size (preferably above 400, Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021). These expectations were matched for all effect sizes with few exceptions. These exceptions occurred when the data were insufficient to estimate uncertainty about some effects, such as case-level data with 100% NAP estimate based on sample statistics.

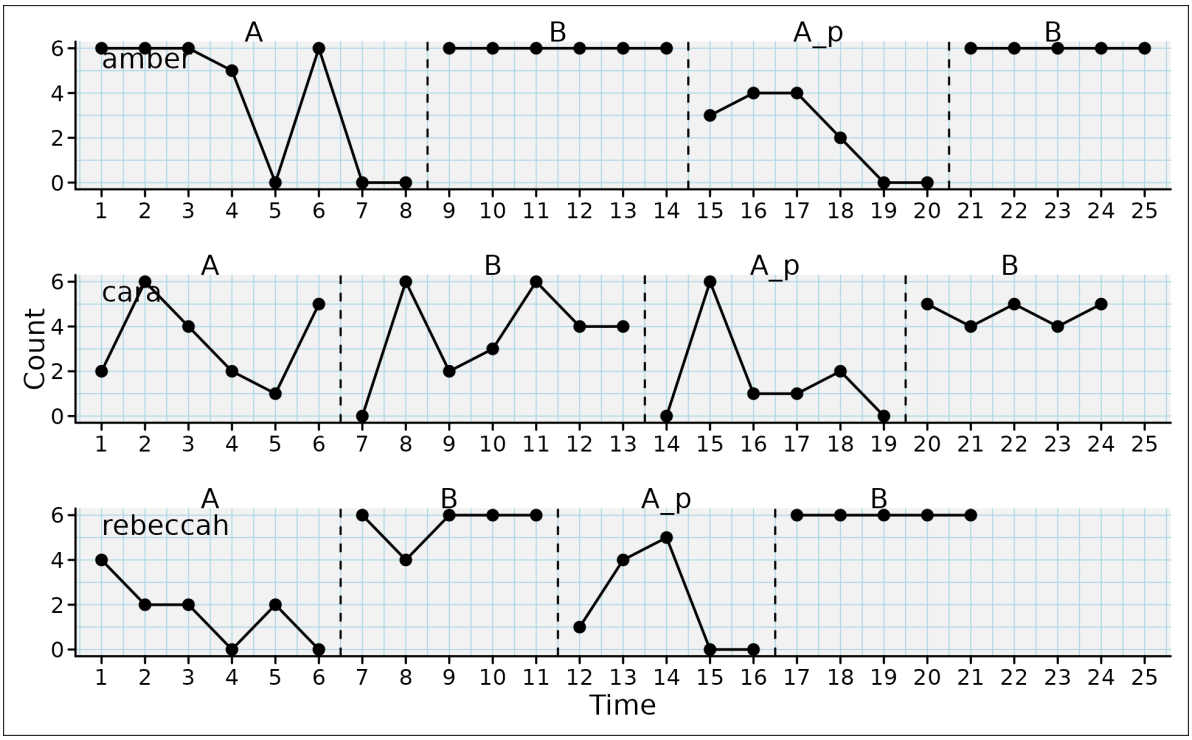
For the Tasky et al. (2008) study, we examine model fit using posterior predictive checking (Gelman, Meng, & Stern, 1996; Rubin, 1984). Precisely, we ask: how well does data generated from the model (or the model's claims about the data) reflect the observed data? Since we use Bayesian estimation to augment the information in the data, we expect that the model's claims about the data should reflect the observed data. We make this assessment by comparing the distribution of the data according to the model to the distribution of the data.

After fitting the models, we compute effect sizes based on the model. We focus on the NAP for Shogren et al. (2004) study. Since outcomes in the Tasky et al. (2008) study can be expressed as proportions, we compute the log odds and rate ratios, in addition to non-overlap effect sizes, the NAP and PEM. We also compare the estimates to those obtained using standard effect size computation methods, as obtained from the *SingleCaseES* package. This means the log rate and odds ratios included small-sample

bias-corrections (Pustejovsky, 2015). The NAP statistic is based on counting pairwise comparisons, with confidence intervals based on the method by (Newcombe, 2006, method 5). The PEM is similarly counting-based, but its sampling distribution is yet to be described. We refer to these standard effect sizes as *raw* effect sizes. Similar to effect sizes returned by the HOM, the raw effect sizes assume no time trend in the data. Unlike effect sizes returned by the HOM, inferences about the raw effect sizes assume independent data points within each phase. All data demonstration details, including R code, are reported in the *demo.html* file in the online OSF repository.

Figure 1

Tasky et al. (2008): Time series



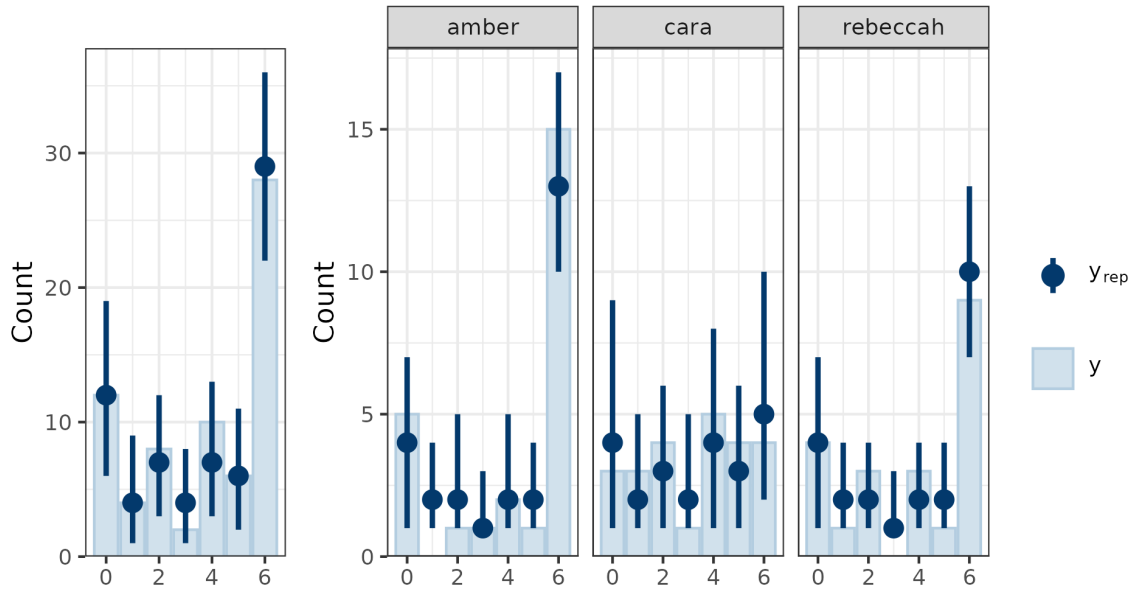
Note. *B* phases are intervention periods. The goal was to increase the number of times out of 6 that each participant stayed on-task.

Tasky et al. (2008)

The Tasky et al. (2008) time series are shown in Figure 1. A casual visual inspection suggests that the intervention is successful at increasing on-task behavior for Amber and Rebeccah – the results are less clear for Cara. We fit the HOM and assessed

model fit using posterior predictive checking – see Figure 2. The frequency of each observed response and the expectations based on the model for the overall sample and by case are generally aligned, suggesting that the model reflects patterns in the data. The degree of autocorrelation for these data was uncertain and likely small, estimate = -0.04, 95% credible interval (CrI) [-.39, .31].

Figure 2
Tasky et al. (2008): Posterior predictive checking

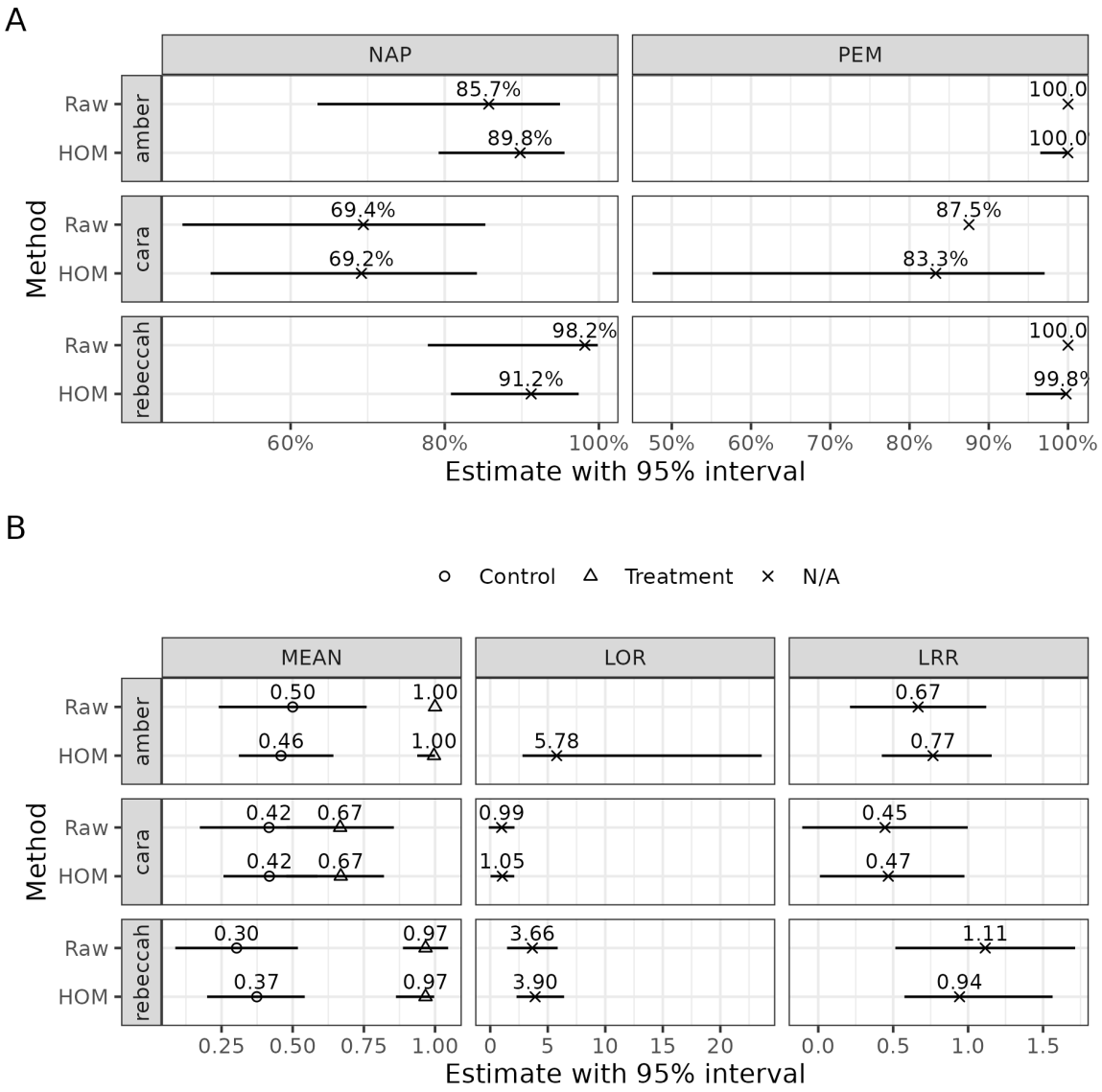


Note. Bar charts show the frequency of each count. Error bars show the expected range of each count based on the model. Both raw data and model-based expectations should be aligned.

We now turn to estimation of non-overlap effect sizes, reported in panel A of Figure 3. The HOM NAP estimates for Amber and Rebeccah are very large (about 90%) and fairly certain, matching our casual visual inspections. Cara's NAP, while large on average (about 70%), is much less certain with a 95% CrI extending from about 50% to 85%. The results using standard methods largely agree with the HOM, with one noticeable difference: standard methods result in wider intervals than HOM intervals. This happens because the HOM borrows information across cases. For the PEM, the HOM is able to return intervals, and these results tell a similar story to the NAP-based patterns.

Figure 3

Tasky et al. (2008): Effect sizes estimates with 95% intervals



Note. Panel A. Non-overlap effect sizes. Panel B. Location-based effect sizes. There are no *raw* intervals for the PEM as its sampling distribution is not yet described. Hierarchical ordinal model (*HOM*) point estimates are posterior medians. All of Amber's treatment data points were 100%. Accordingly, there is no uncertainty about her treatment sample mean, and her sample log odds ratio is undefined.

The first location-based effect size we report is the mean of each case in each phase, see panel B of Figure 3. The HOM means have narrower intervals than the t-distribution intervals for the sample means. All of Amber's treatment data points were 100%. Accordingly, there is no uncertainty about her treatment sample mean. Amber's raw log odds ratio is also undefined. However, since the HOM borrows information across cases and places a weakly-informative prior on coefficients,[4] the HOM estimates Amber's log odds ratio to be both very large and highly uncertain.

The raw log odds ratios for other cases are slightly smaller and have narrower intervals than the HOM log odds ratios. This is due to the bias correction applied to these estimates. Finally, the log rate ratios returned by the HOM have narrower intervals than the raw log rate ratios, likely due to borrowing information across cases.

Shogren et al. (2004)

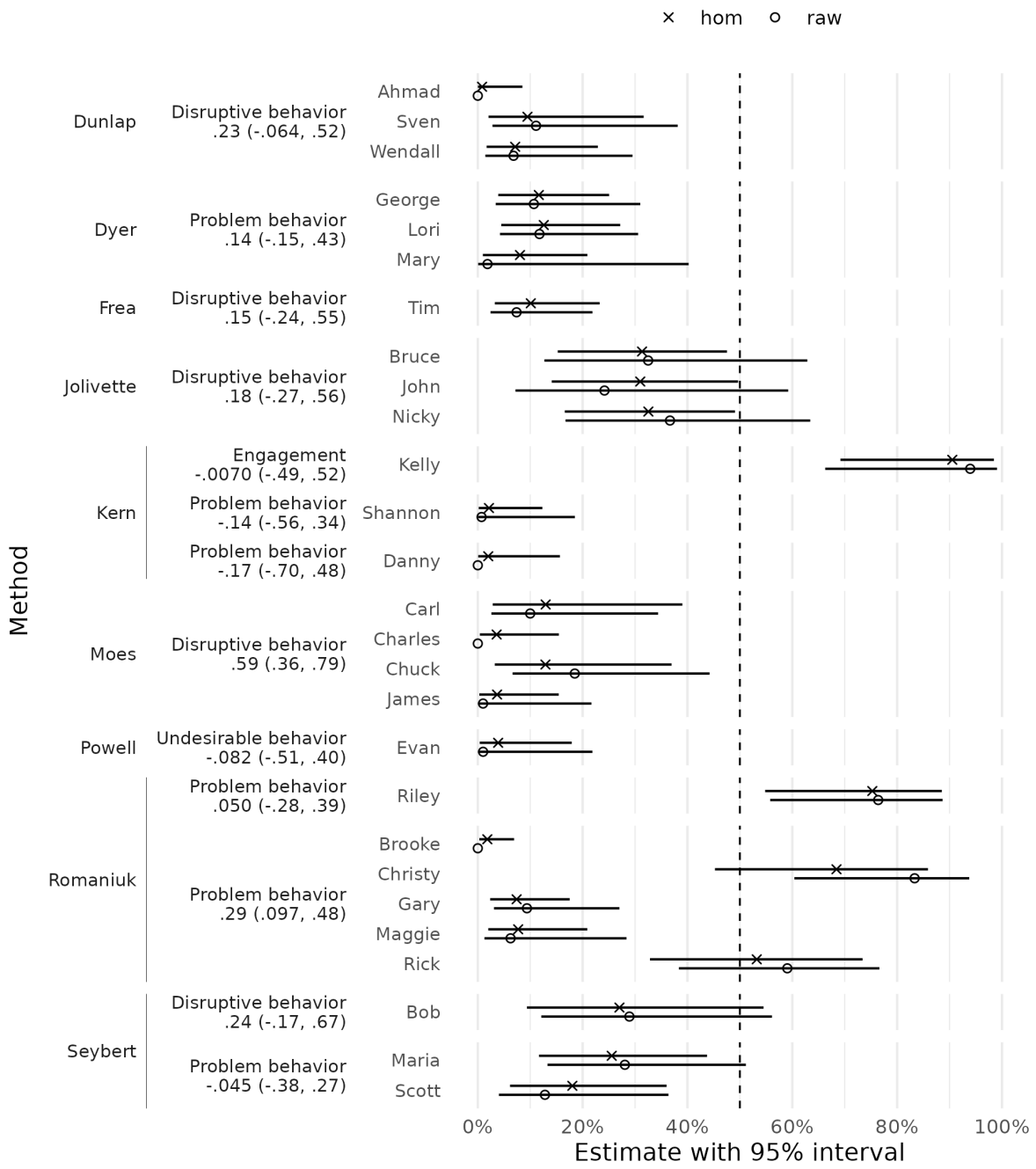
For these data, we fit a separate model for each study, outcome measure, measurement scale (percentages, counts, or rates), and effect direction (increase or decrease desired), leading to 13 different models across the nine constituent studies. This demonstration allows us to observe the estimation of effect sizes across various problems. For example, the analysis of *disruptive* behaviors in the study by Frea, Arnold, and Vittimberga (2001), all participants in Kern, Mantegna, Vorndran, Bailin, and Hilt (2001), *undesirable* behaviors in Powell and Nelson (1997), Riley's *problem* behaviors in Romaniuk et al. (2002) and *disruptive* behaviors in Seybert, Dunlap, and Ferro (1996) were analyses with a single participant, so there is no borrowing of information across participants. At the other end, five of six cases in the Romaniuk study are in a single model, enabling information borrowing across cases. Case-wise HOM and raw NAP estimates are presented in Figure 4.

---

[4] In the situation that all of the data are Amber's and there was no other case to borrow information from, the treatment effect in equation 4 would reduce to β_0. This coefficient has a $\mathcal{N}(0, 2.5)$ prior, such that Amber would still have a log odds ratio based on the constraints defined by the prior.

Figure 4

Shogren et al. (2004): Effect sizes estimates with 95% intervals



Note. We show the autocorrelation coefficient with its 95% interval under the measures. Hierarchical ordinal model (*HOM*) point estimates are posterior medians.

Behavior of NAP estimates for multiple participant analyses. For analyses with multiple participants, we see the same patterns identified in the Tasky et al. (2008) results: HOM estimates are more similar across participants and have narrower intervals than the raw NAP estimates. However, when a participant is highly dissimilar from others (e.g., Christy in the Romaniuk et al. (2002) study), their HOM NAP intervals are not narrower than the raw NAP intervals. In Christy's case, the HOM NAP intervals communicate skepticism that her NAP estimate is above 50% as compared to the raw NAP intervals whose 95% interval is clearly above 50%. We show the time series for these five Romaniuk et al. (2002) study cases in Figure A1 in appendix A. The supplementary demonstration file (*demo.html*) also includes checks for Christy, which suggest the HOM reasonably captures her data.

Behavior of NAP estimates for single participant analyses. The behavior of HOM NAP estimates relative to raw NAP estimates depends on how extreme the NAP estimate is. When the NAP estimate is extreme (e.g., Shannon and Danny in Kern et al. (2001) or Evan in Powell and Nelson (1997)), the HOM estimates are less extreme than the raw estimates and have narrower intervals. This is the result of the weakly-informative prior on the treatment effect in the HOM, $\beta_0 \sim \mathcal{N}(0, 2.5)$. When the NAP estimates are not extreme (e.g., Riley in Romaniuk et al. (2002) or Bob in Seybert et al. (1996)), the HOM estimates produce similar intervals to the raw estimates since the weakly-informative prior on the treatment effect has less of an impact.

Simulation studies

We conducted two simulation studies to examine the HOM's ability to estimate several effect sizes for two common response scales in single-case research: unbounded counts in Study 1 and percentages (often obtained from bounded counts) in Study 2. For both studies, we are interested in the accuracy of effect size estimates at the level of each case within a study. We focus on simple treatment-reversal designs. For studies with this design to be eligible to meet the What Works Clearinghouse *standards with reservations*,

there must be two phases per condition with at least three data points per phase (What Works Clearinghouse, 2022, pg. 113). Accordingly, we simulate data according to an ABAB design, with at least three data points per phase, for all studies.

Study 1 – Unbounded counts

Data generation process

The observed counts, y_{ct}, for case $c \in \{1, \ldots, C\}$ at time $t \in \{1, 2, \ldots, 4 \cdot k\}$ where k is phase-length were generated according to the following DGP:

$$\mathbf{z}_c \sim \mathcal{N}_T(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}),\ \mu_{ct} = \gamma_c + \beta_c \cdot x_{ct},\ \sigma_{ij} = \tau_1^2 \cdot \rho^{|i-j|} \text{ for } i, j \in \{1, \ldots, T\}, \tag{5}$$

$$x_{ct} = \left\lfloor \frac{t-1}{k} \right\rfloor \mod 2, \tag{6}$$

$$\gamma_c \sim \mathcal{N}(\gamma_0, \tau_2),\ \beta_c \sim \mathcal{N}(\beta_0, \tau_2), \tag{7}$$

$$y_{ct} \sim \text{Poisson}\left(\exp(z_{ct})\right) \tag{8}$$

where $x_{\cdot\cdot}$ is an indicator for a treatment phase measurement. Given this DGP, the count vector for each case, $\mathbf{y}_c$, is a multivariate Poisson log-normal variable (Aitchison & Ho, 1989). The case-specific mean on the latent scale is dependent on a hierarchical model, with separate random-effects on both the case baseline level (γ_c) and the case treatment effect (β_c). The latent multivariate normal variable has a scale of τ_1 and an autocorrelation of ρ.

Given this DGP, the means, variances, and correlations for the observed counts are (see equations 2.3 – 2.6 in Aitchison & Ho, 1989):

$$\mathbb{E}\left(y_{ct}\right) = \exp\left(\mu_{ct} + 0.5 \cdot \sigma_{tt}\right) = \alpha_{ct}, \tag{9}$$

$$\text{Var}\left(y_{ct}\right) = \alpha_{ct} + \alpha_{ct}^2\left(\exp(\sigma_{tt}) - 1\right) \tag{10}$$

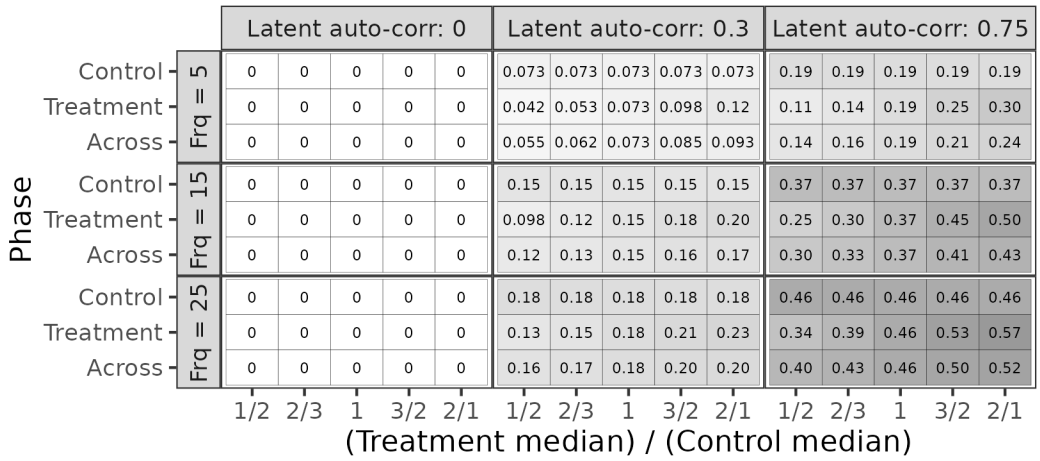
$$\text{Corr}(y_i, y_j) = \frac{\exp(\sigma_{ij}) - 1}{\left[\left(\exp(\sigma_{ii}) - 1 + \alpha_i^{-1}\right)\left(\exp(\sigma_{jj}) - 1 + \alpha_j^{-1}\right)\right]^{\frac{1}{2}}} \tag{11}$$

Data generated under this process will necessarily be overdispersed since $\text{Var}(y_{ct}) > \alpha_{ct}$. Additionally, the observed autocorrelation will have a lower magnitude than the latent

correlation.

Figure 5

Autocorrelation on count scale



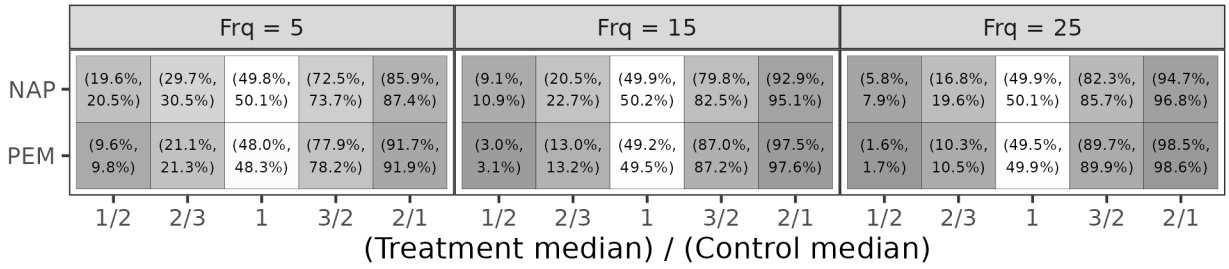
Note. Cells with larger correlations are darkened. Correlations vary as a function of latent autocorrelation and the means of the pair of correlated data points. For *Across*, one data point is in the control phase with the other in the treatment phase.

We varied several values in the DGP above. There were three levels of $\exp(\gamma_0) \in \{5, 15, 25\}$ matching data with median baseline counts of 5, 15. and 25 respectively across all cases. This covers the range of baseline frequencies commonly found in single-case studies, based on a 2023 review of measurement procedures in single-case designs (Pustejovsky, Swan, & English, 2023). We had five levels of $\exp(\beta_0) \in \{1/2, 2/3, 1, 3/2, 2\}$ reflecting treatment effects ranging from halving the median baseline level to doubling the median baseline level across all cases. We varied the degree of latent autocorrelation from none to high autocorrelation, $\rho \in \{0, .3, .75\}$ – see Figure 5 for the observed scale correlations averaged over cases based on equation 11. We also varied the number of cases, $C \in \{2, 3, 4\}$. We fixed the latent multivariate normal scale parameter, τ_1, to 0.25, a small but significant value such that the observed counts were never extremely high. We also fixed the random-effect scale for the hierarchical models on the latent mean coefficients, τ_2, to 0.15 for similar reasons. The ABAB design had three phase length levels, $k \in \{3, 5, 10\}$ measurements per phase. In total, there were 405 design

conditions ($3\gamma_0 \cdot 5\beta_0 \cdot 3\rho \cdot 3C \cdot 3k$). Finally, we note that the simulation study by Swan and Pustejovsky (2018) served as a guide for the levels of some of the design factors above.

Figure 6

Expected values of non-overlap effect sizes by simulation condition



Note. Cells with larger effects are darkened. The expected effects depend on the autocorrelation and phase-length, so we show the minimum and maximum expected values. The range of the NAP is greater than the range of the PEM.

Effect sizes

We are interested in estimating the case-wise treatment effects. In addition to the non-overlap indices, we also track the case-wise log rate ratio (LRR). The LRR for case c is β_c. We obtained the average NAP and PEM given γ_0, β_0, and τ_1 (fixed at 0.25) using Monte Carlo methods (the OSF repository contains code to implement these methods); these statistics varied by phase-length and autocorrelation. These effect sizes are reported in Figure 6.

We used the proposed ordinal model to estimate the three effects for each case. For each model, we requested 1000 posterior samples across three chains and discarded the first 500 samples per chain. This left 1500 (500×3 chains) posterior samples per case-wise effect, on which we based inference. For the point estimate, we relied on the posterior mean. For uncertainty quantification, we relied on the 90% quantile interval. We did not use the 95% interval because a more extreme interval may require more posterior samples for stable estimation.

For our business-as-usual or standard methods, we estimated all three effects as implemented in the *SingleCaseES* package.

Outcomes

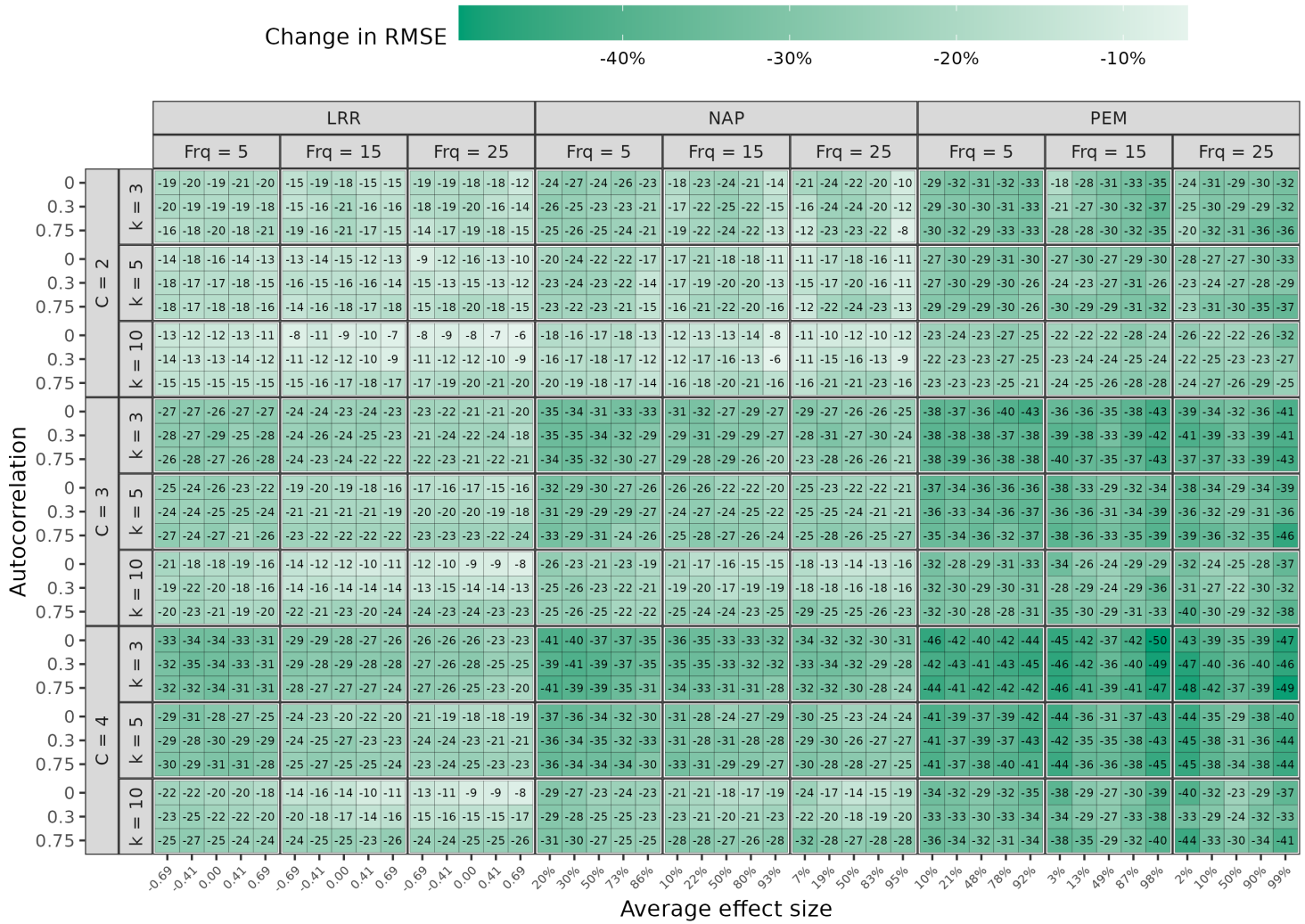
Let R be the number of simulation iterations per condition, we set $R = 500$.

Accuracy. We were interested in the accuracy of the proposed method compared to the standard approach for computing the three effects above, so we computed the root mean squared error $\left(\text{RMSE} = \sqrt{\frac{1}{R}\sum_{i=1}^{R}\left[\frac{1}{C}\sum_{c=1}^{C}\left(\hat{\theta}_{ic} - \theta_{ic}\right)^2\right]}\right)$ to compare the proposed model to the methods based on sample statistics, where $\hat{\theta}_{ic}$ was the point estimate for some effect, θ, for case c in simulation iteration i. Then we computed the percentage change in RMSE with the RMSE based on sample statistics as a baseline. We expect the proposed method to be relatively more accurate when there are more cases, as there will be more borrowing of information. We also expect the difference in accuracy to be reduced when the phase length is longer, as the sample statistics will be more stable due to having more data available.

Coverage of the 90% interval. We were interested in the ability of the proposed method to maintain adequate coverage, so we computed the percentage of times the 90% quantile interval included the true parameter at the level of each case-wise effect. We hope this value is 90% for effects from the proposed model and deem coverage in (87.5%, 92.5%) and (85%, 95%) as ideal and acceptable, respectively. We expect the sample statistic based effects to have less than adequate coverage when the latent autocorrelation is non-zero.

Figure 7

Simulation Study 1: Relative accuracy of HOM effect sizes



Note. % drop in RMSE is reported in each cell, and cells with lower RMSEs are darker. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 6 for population parameters.

Results

We present results as Figures; see study_1_tables.xlsx in the OSF repository for the results in tables.

Accuracy. The HOM effect sizes were more accurate than standard effect sizes across all conditions, see Figure 7. For the LRR, the drop in RMSE ranged between 5% and 34%. For the NAP, the drop in RMSE ranged between 7% and 41%. For the PEM, the drop in RMSE ranged between 18% and 50%. This result is to be expected given the benefits of borrowing information across cases. For each effect, the relative accuracy improves with more cases. And given a number of cases, the relative accuracy decreases with increased phase length.

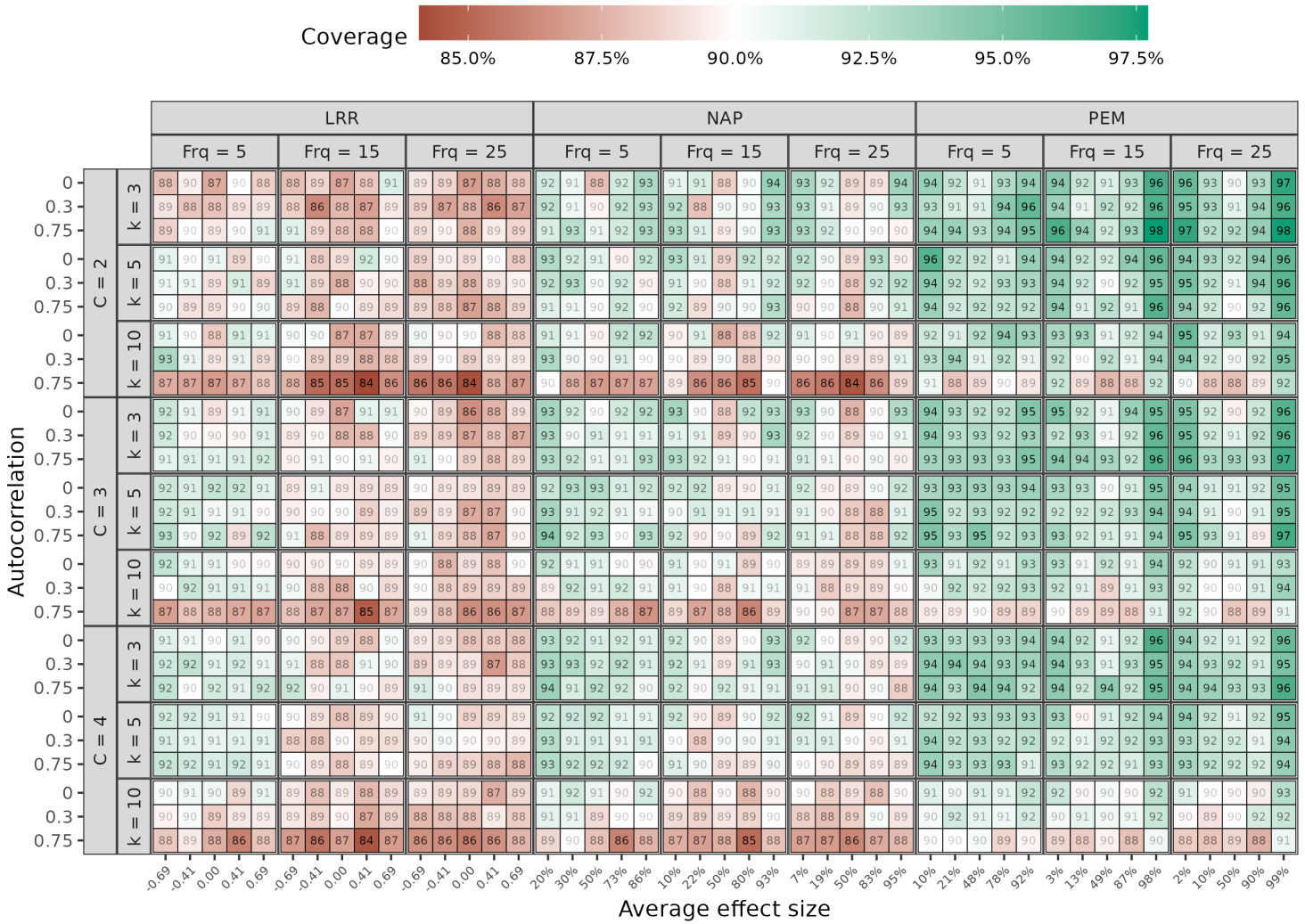
Coverage. The HOM effect sizes often had acceptable coverage, i.e., in (85%, 95%) (see Figure 8). Coverage was very rarely less than 85% and sometimes more than 95%. Coverage was more than 95% when the effect was very extreme (e.g., the smallest and largest PEMs when the baseline frequency was high) and the phase length was short. Coverage was more likely to be low given the combination of long phase length and high latent autocorrelation. This was especially common for the LRR.

Additional HOM results. We also examined bias of the HOM estimates, reported in Figure B2 in appendix B. The HOM estimates were somewhat biased towards the null, such that estimates were often less extreme relative to their respective population parameters. We also examined the returned autocorrelation. As noted in the DGP section, the response scale autocorrelation is always less than the latent autocorrelation and depends on phase means, such that the autocorrelation should be estimated separately by phase. However, the HOM, as developed, estimates a single autocorrelation parameter, so we are unable to compute autocorrelation bias. Hence, we report the bias-corrected coverage of the autocorrelation parameter. The coverage was often adequate except under high latent autocorrelation coupled with non-short phase-lengths, see Figure B1 in appendix B. This pattern arises because the same latent autocorrelation yields different

autocorrelations for the control and treatment phases on the count scale. This difference is greater for larger latent autocorrelation (see Figure 5), resulting in suboptimal HOM autocorrelation recovery. The problem becomes more severe at longer phase lengths because data from longer phases provide more evidence against the assumption of identical autocorrelations between the control and treatment phases on the count scale.

Figure 8

Simulation Study 1: Empirical coverage of 90% uncertainty intervals of HOM effect sizes



Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 6 for population parameters.

Results for sample statistics based effect sizes. These methods analyze data separately for each case, so the number of cases is not a factor in performance. The bias in these estimates was ignorable, see Figure B3. However, the returned intervals often demonstrated severe undercoverage, see Figure B4. Undercoverage was more severe for both the LRR and NAP as the latent correlation increased. Undercoverage was also a problem for extreme values of NAP, regardless of the latent autocorrelation.

Study 2 – Bounded counts

Study 2 shared several elements with Study 1. The data generation process for data in this study shared equations 5 and 7 for simulating latent normal variables. However, the observed counts, y_{ct} were assumed to be binomial, $y_{ct} \sim \text{Binomial}\left(10, [1 + \exp(-z_{ct})]^{-1}\right)$, where 10 is the number of trials, and the probability of success was a logit-normal variable. Hence, the distribution of observed counts was multivariate binomial logit-normal (Coull & Agresti, 2000).

We varied several values in this DGP. There were three levels of $\exp(\gamma_0) \in \{10\%, 20\%, 40\%\}$ matching data with median baseline probability of success of low to average across all cases. This approximately matches the range of baseline proportions common in single-case studies where an increase in this proportion is desired based on a 2023 review of measurement procedures in single-case designs (Pustejovsky, Swan, & English, 2023). Unlike Study 1, we did not consider the reverse where a decrease in the proportion is desired since the binomial logit-normal distribution behaves similarly at low and high probabilities. We had three levels of $\exp(\beta_0) \in \{1, 3/2, 2\}$ reflecting treatment effects ranging from null to doubling the median odds of the baseline level on average across all cases. We varied the degree of latent autocorrelation from none to high autocorrelation, $\rho \in \{0, .3, .75\}$. Unlike the multivariate Poisson log-normal, the moments of the multivariate binomial logit-normal do not have an analytic solution. We also varied the number of cases, $C \in \{2, 3, 4\}$. We varied the logit-normal scale parameter, $\tau_1 \in \{0.5, 1\}$ to reflect varying levels of overdispersion. Unlike Study 1, increases in τ_1 will

not result in extremely high counts. We fixed the random-effect scale for the hierarchical models on the latent mean coefficients, τ_2, to 0.15 to keep the number of design conditions manageable. As with Study 1, we chose an ABAB design with three phase length levels, $k \in \{3, 5, 10\}$ measurements per phase. In total, there were 486 design conditions ($3\gamma_0 \cdot 3\beta_0 \cdot 3\rho \cdot 3C \cdot 2\tau_1 \cdot 3k$).

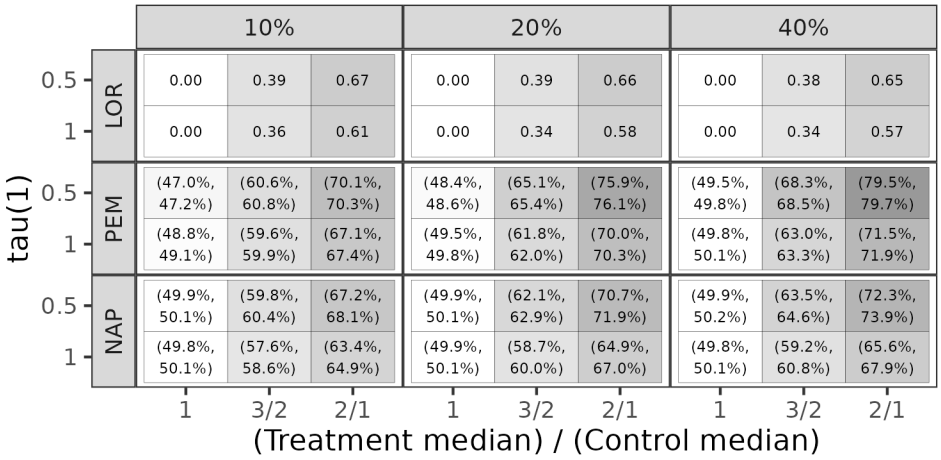
We repeated each design condition 500 times, as in Simulation Study 1, and maintained accuracy and coverage as primary outcomes.

Effect sizes

We are interested in estimating the case-wise treatment effects. In addition to the non-overlap indices, we also track the case-wise log odds ratio (LOR). The LOR for case c is β_c. We obtained the average LOR, NAP and PEM given γ_0, β_0 and τ_1 using Monte Carlo methods (available in the OSF repository). These effect sizes are reported in Figure 9.

Figure 9

Expected values of effect sizes by simulation condition

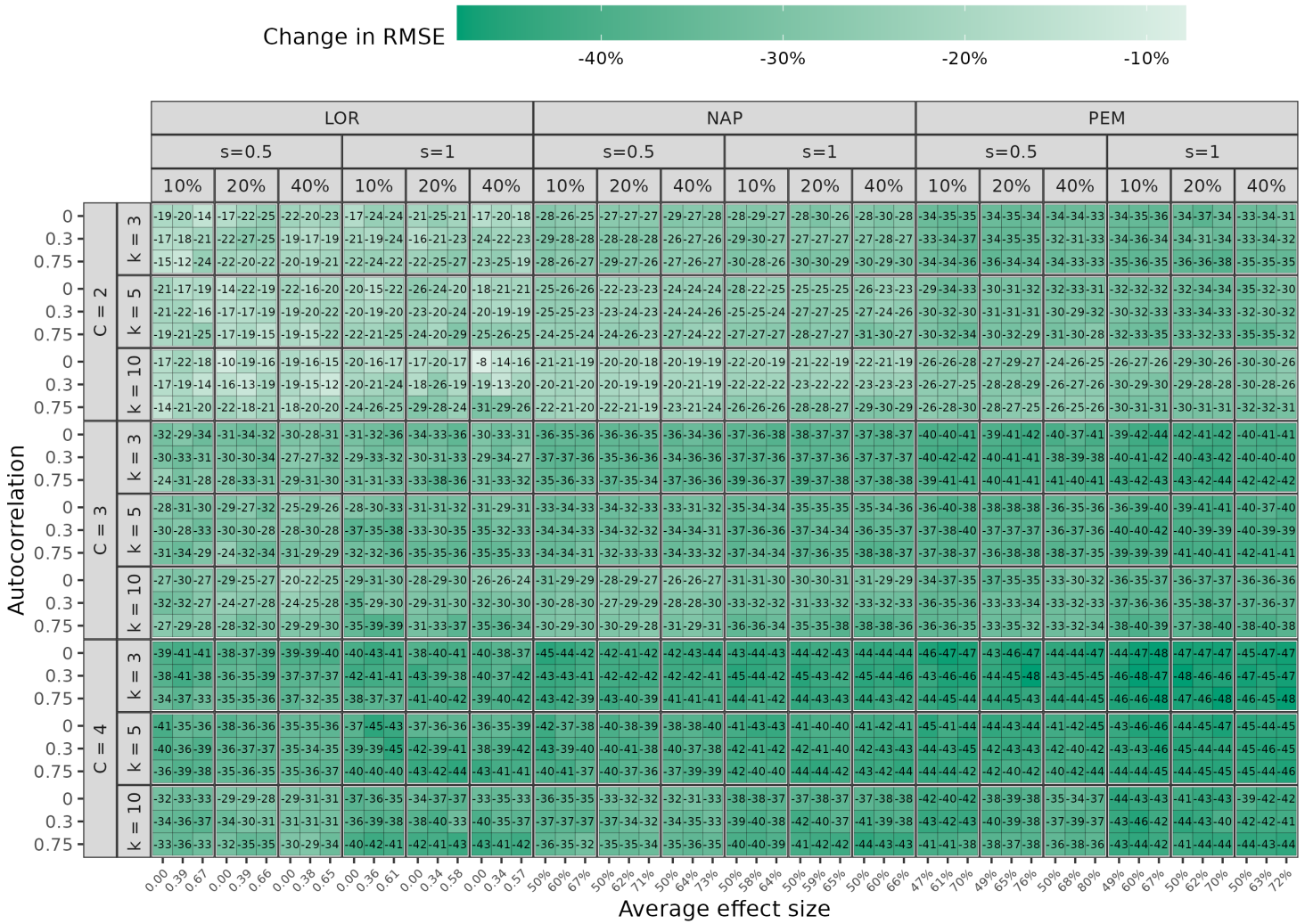


Note. Cells with larger effects are darkened. The expected PEM and NAP depend on the autocorrelation and phase-length, so we show the minimum and maximum expected values. The range of the NAP is greater than the range of the PEM.

We maintained the same MCMC estimation settings for estimating the effect sizes as in Simulation Study 1, and also estimated the sample statistics based effects using the *SingleCaseES* package.

Figure 10

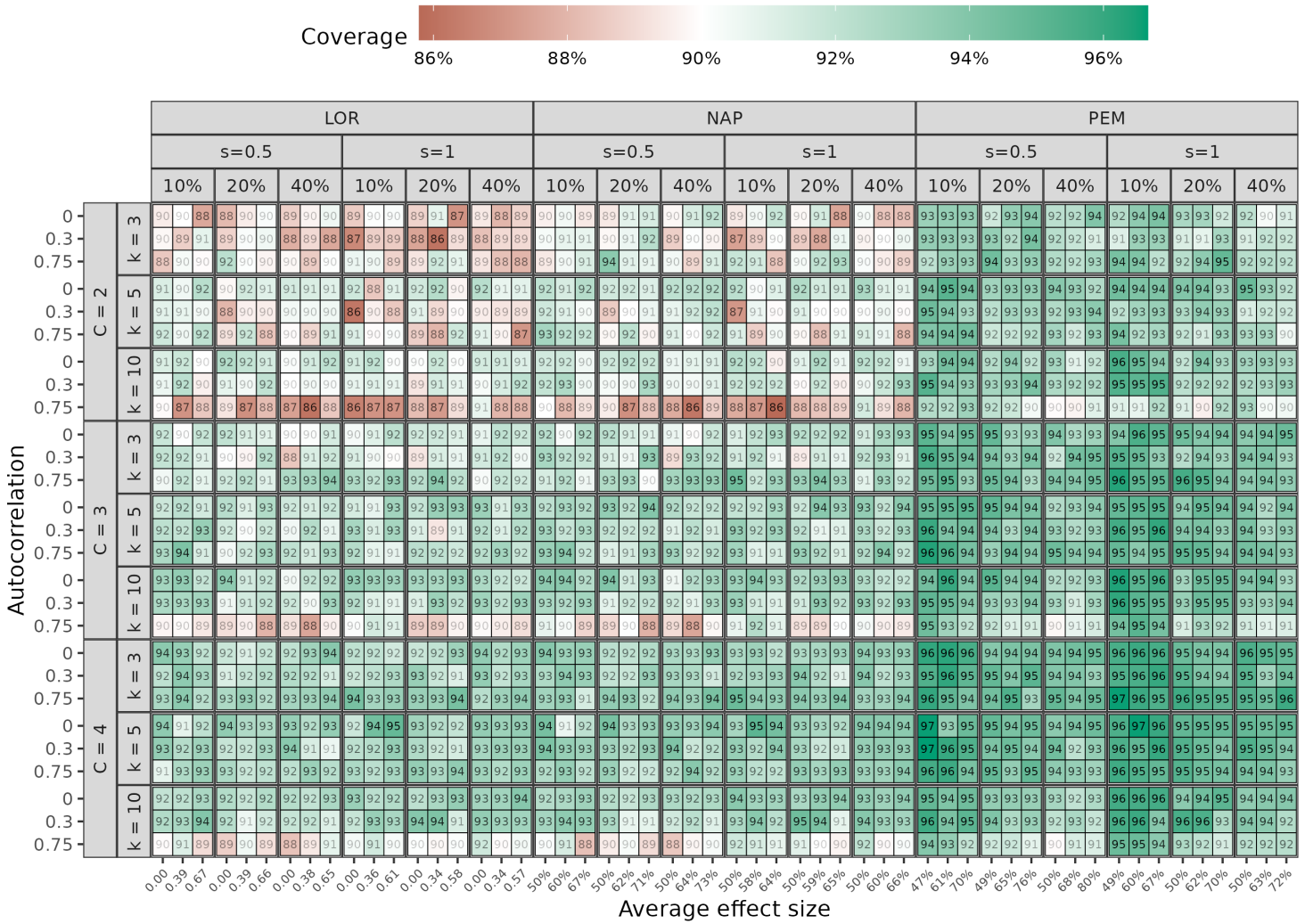
Simulation Study 2: Relative accuracy of HOM effect sizes



Note. % drop in RMSE is reported in each cell, and cells with lower RMSEs are darker. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 9 for population parameters.

Figure 11

Simulation Study 2: Empirical coverage of 90% uncertainty intervals of HOM effect sizes



Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 9 for population parameters.

Results

The patterns of results were similar to Study 1, with somewhat more desirable performance for the HOM. We note that the range of effects in this study did not include as extreme effects as there were in Study 1. We present results as Figures; see study_2_tables.xlsx in the OSF repository for the results in tables.

Accuracy. We found the same pattern of results as in Study 1: the HOM effect sizes were more accurate than standard effect sizes across all conditions, see Figure 10. For the LOR, the drop in RMSE ranged between 8% and 45%. For the NAP and PEM, the drop in RMSE ranged between 18% and 48%.

Coverage. The HOM effect sizes often had acceptable coverage, i.e., in (85%, 95%) (see Figure 11). Coverage was never less 86%, so under-coverage was not a problem. However, coverage was often more than 95% for the PEM.

Additional HOM results. We also examined the bias of the HOM estimates, reported in Figure C2 in appendix C. The HOM estimates were sometimes notably biased. With regard to bias-corrected coverage of the autocorrelation parameter, the coverage was often adequate except under high latent autocorrelation coupled with non-short phase-lengths, see Figure C1 in appendix C.

Results for sample statistics based effect sizes. The bias in these estimates was negligible, see Figure C3. Additionally, the returned intervals had adequate coverage when the latent autocorrelation was zero, see Figure C4. As would be expected, undercoverage became a problem as the latent autocorrelation increased.

Discussion

We have presented a method for computing case-wise non-overlap and *parametric* effect sizes for SCD data with treatment-reversal designs. We have demonstrated the method with real-world datasets and shown its desirable performance via simulation. This work addresses key limitations in existing SCD effect sizes, including the inability of traditional non-overlap indices to account for autocorrelation, provide reliable uncertainty

quantification, or borrow information across cases. The proposed HOM offers enhanced precision and robustness in SCD effect size estimation.

Although we have focused on computing non-overlap effect sizes, we note that the method is much more general, with potential as a general model for analysis of SCD data. Potential extensions of the method include computing design-comparable effect sizes (Pustejovsky, Hedges, & Shadish, 2014) or conducting meta-analysis of multiple single-case studies. The current model has all the elements needed to compute an estimate of the between-case standardized mean difference, assuming no time trend. However, additional random effects would be needed to capture differences across studies for meta-analyses. Exploring these extensions to facilitate such analyses would bridge the gap between SCD and broader evidence synthesis, enabling researchers to contribute to the growing body of meta-analytic research in the behavioral and educational sciences. We intend to explore these extensions in future studies.

SCD data often have limited samples, making it difficult to verify or critique distributional assumptions. If SCD data are assumed unidimensional, then they are also ordinal. Hence, the assumption of ordinality (while accommodating autocorrelation) is an appealing aspect of the proposed method, as it respects the discrete nature of behavioral measurements

The practical implications of this method are significant for single-case researchers. The HOM enables researchers to derive effect sizes that are not only more precise but also accompanied by uncertainty intervals, addressing the long-standing issue of uncertainty quantification in SCD analysis. Additionally, the Bayesian framework allows for regularization and the incorporation of prior knowledge, which can be especially useful in studies with short phase lengths or low variability. These features are critical for practitioners seeking to make informed decisions about intervention effectiveness, particularly for low-incidence populations where data points may be few.

Inherent to the model as developed is the assumption that different cases within the

same study are similar, such that the results of one case should inform what we learn about other cases. As shown in both simulation studies, this assumption built into the model can improve the accuracy of case-specific estimates. Within a single-case study, a researcher could obtain standard effect sizes for a case that conflict with the effect sizes from the HOM approach. This can occur for two reasons: (i) the single case is markedly different from other cases; (ii) the case's data are not informative, possibly due to short phase lengths. The researcher may then wonder if the HOM estimates for the case are accurate. In such situations, we recommend case-wise posterior predictive checking as shown in Figure 2. When the HOM reasonably captures the data for a case, we believe the estimates for the case may be trustworthy.

One of the assumptions of the model, as developed, is that time specifically affects the outcome via an autocorrelation process. One could relax this assumption by estimating a more general correlation structure, such as via latent case-specific Gaussian processes (Rasmussen & Williams, 2005). Gaussian processes are appealing due to their adaptability and ability to model complex data. This could enhance the model's flexibility, allowing it to accommodate more complex behavioral data patterns.

Another development would be better-developed model criticism. In the demonstration with data from the Tasky et al. (2008) study, we used posterior predictive checks of the frequency of each observed response to assess the model's capacity to reflect patterns in the data. However, the limited amount of data makes it difficult to extensively criticize the model. Future work should focus on developing methods that allow researchers to systematically evaluate model fit, even in the presence of limited data.

In summary, the HOM introduced in this study offers a robust solution for analyzing SCD data and addresses a longstanding gap in the statistical infrastructure for non-overlap effect sizes. By providing a framework that explicitly accounts for autocorrelation, uncertainty, and ordinality of behavioral data, the model represents a significant advancement in the field. The development of the accompanying R package

increases accessibility, enabling researchers to readily apply the method to their data. Future research should focus on expanding the model to accommodate additional SCD designs (e.g., multiple-baseline or alternating treatments), explore integration with meta-analytic frameworks, and investigate methods for modeling complex temporal dynamics. These directions will help ensure that the HOM continues to evolve as a foundational tool for rigorous, interpretable, and flexible analysis of SCD research.

References

Aitchison, J., & Ho, C. H. (1989, December). The multivariate Poisson-log normal distribution. *Biometrika*, *76*(4), 643–653. doi: 10.1093/biomet/76.4.643

Allison, D. B., & Gorman, B. S. (1994, November). "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

Barnard-Brak, L., Watkins, L., & Richman, D. M. (2021). Autocorrelation and estimates of treatment effect size for single-case experimental design data. *Behavioral Interventions*, *36*(3), 595–605. doi: 10.1002/bin.1783

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle: A review, generalizations, and statistical implications* (2nd ed., Vol. 6). Institute of Mathematical Statistics.

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*(2), 332–340. doi: 10.1037/a0034745

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. doi: 10.1037/0033-2909.114.3.494

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge. doi: 10.4324/9780203771587

Coull, B. A., & Agresti, A. (2000, March). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, *56*(1), 73–80. doi: 10.1111/j.0006-341X.2000.00073.x

Frea, W. D., Arnold, C. L., & Vittimberga, G. L. (2001, October). A demonstration of the effects of augmentative communication on the extreme aggressive behavior of a child with autism within an integrated preschool setting. *Journal of Positive Behavior Interventions*, *3*(4), 194–198. doi: 10.1177/109830070100300401

Gast, D. L., & Ledford, J. R. (Eds.). (2014). *Single case research methodology: Applications in special education and behavioral sciences, 2nd ed.* New York, NY, US:

Routledge/Taylor & Francis Group. doi: 10.4324/9780203521892

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760. Retrieved from http://www.jstor.org/stable/24306036

Harrison, X. A. (2014, October). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616. doi: 10.7717/peerj.616

Harrison, X. A. (2015, July). A comparison of observation-level random effect and beta-binomial models for modelling overdispersion in binomial data in ecology & evolution. *PeerJ*, *3*, e1114. doi: 10.7717/peerj.1114

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research synthesis methods*, *4*(4). doi: 10.1002/jrsm.1086

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005, January). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179. doi: 10.1177/001440290507100203

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290.

Kern, L., Mantegna, M. E., Vorndran, C. M., Bailin, D., & Hilt, A. (2001, January). Choice of task sequence to reduce problem behaviors. *Journal of Positive Behavior Interventions*, *3*(1), 3–10. doi: 10.1177/109830070100300102

Kotz, S., Lumelskii, Y., & Pensky, M. (2003). *The stress-strength model and its generalizations*. World Scientific. doi: 10.1142/5015

Kowal, D. R., & Wu, B. (2023, June). Semiparametric count data regression for self-reported mental health. *Biometrics*, *79*(2), 1520–1533. doi: 10.1111/biom.13617

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf,

D. M., & Shadish, W. R. (2013, January). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. doi: 10.1177/0741932512452794

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. doi: 10.1111/oik.05985

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods*, *28*(3), 719–739. doi: 10.1037/met0000443

Li, J. C.-H. (2015, October). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, *48*(4), 1560–1574. doi: 10.3758/S13428-015-0667-Z

Liu, Q., Shepherd, B. E., Li, C., & Harrell, F. E. (2017, November). Modeling continuous response variables using ordinal regression. *Statistics in medicine*, *36*(27), 4316–4335. doi: 10.1002/sim.7433

Ma, H.-H. (2006, September). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification*, *30*(5), 598–617. doi: 10.1177/0145445504272974

MacKinnon, J., & White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(537), 305–325. (Publisher: North-Holland) doi: 10.1016/0304-4076(85)90158-7

Maggin, D. M., Cook, B. G., & Cook, L. (2019, August). Making sense of single–case design effect sizes. *Learning Disabilities Research & Practice*, *34*(3), 124–132. doi: 10.1111/ldrp.12204

Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017, November). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education*, *38*(6), 323–330. doi: 10.1177/0741932517717043

Maggin, D. M., & Odom, S. L. (2014, April). Evaluating single-case research data for

systematic review: A commentary for the special issue. *Journal of School Psychology*, *52*(2), 237–241. doi: 10.1016/j.jsp.2014.01.002

Manuguerra, M., & Heller, G. Z. (2010, April). Ordinal regression models for continuous scales. *The International Journal of Biostatistics*, *6*(1). doi: 10.2202/1557-4679.1230

Manuguerra, M., Heller, G. Z., & Ma, J. (2020, December). Continuous ordinal regression for analysis of visual analogue scales: The r package ordinalCont. *Journal of Statistical Software*, *96*(1), 1–25. doi: 10.18637/JSS.V096.I08

McCullagh, P. (1980, January). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127. doi: 10.1111/j.2517-6161.1980.tb01109.x

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. doi: 10.1037/0033-2909.111.2.361

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014, April). 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

Newcombe, R. G. (2006, February). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Statistics in medicine*, *25*(4), 559–573. doi: 10.1002/sim.2324

Parker, R. I., & Vannest, K. (2009, December). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–367. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011, July). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303–322. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011, June). Combining nonoverlap and trend for single-case research: Tau-u. *Behavior Therapy*, *42*(2), 284–299. doi: 10.1016/j.beth.2010.08.006

Peterson, B., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *39*(2), 205–217. doi: 10.2307/2347760

Powell, S., & Nelson, B. (1997). Effects of choosing academic assignments on a student with attention deficit hyperactivity disorder. *Journal of Applied Behavior Analysis*, *30*(1), 181–183. doi: 10.1901/jaba.1997.30-181

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, June). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. doi: 10.1016/j.jsp.2018.02.003

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., Chen, M., & Swan, D. M. (2023). SingleCaseES: A calculator for single-case effect sizes [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=SingleCaseES` (R package version 0.7.2)

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014, October). 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., Swan, D. M., & English, K. W. (2023, November). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification*, *47*(6), 1423–1454. doi: 10.1177/0145445519864264

Ramsay, J. O. (1988, November). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441. doi: 10.1214/ss/1177012761

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning* (F. Bach, Ed.). Cambridge, MA, USA: MIT Press.

Rindskopf, D. (2014a). Bayesian analysis of data from single case designs. *Neuropsychological Rehabilitation*, *24*(3-4), 572–589. doi: 10.1080/09602011.2013.866903

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

Romaniuk, C., Miltenberger, R., Conyers, C., Jenner, N., Jurgens, M., & Ringenberg, C. (2002). The influence of activity choice on problem behaviors maintained by escape versus attention. *Journal of Applied Behavior Analysis*, *35*(4), 349–362. doi: 10.1901/jaba.2002.35-349

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172. doi: 10.2307/2240995

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge ; New York: Cambridge University Press.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987, March). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–33. doi: 10.1177/074193258700800206

Seybert, S., Dunlap, G., & Ferro, J. (1996, March). The effects of choice-making on the problem behaviors of high school students with intellectual disabilities. *Journal of Behavioral Education*, *6*(1), 49–65. doi: 10.1007/BF02110477

Shadish, W. R. (2014a, April). 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. (2014b, April). 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., Rindskopf, D. M., & Hedges, L. V. (2008, September). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 188–196. doi: 10.1080/17489530802581603

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014, April). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*(2), 149–178. doi: 10.1016/j.jsp.2013.11.004

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004, October). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, *6*(4), 228–237. doi: 10.1177/10983007040060040401

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014, April). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2), 213–230. doi: 10.1016/j.jsp.2013.12.002

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

Tasky, K. K., Rudrud, E. H., Schulze, K. A., & Rapp, J. T. (2008). Using choice to increase on-task behavior in individuals with traumatic brain injury. *Journal of Applied Behavior Analysis*, *41*(2), 261–265. doi: 10.1901/jaba.2008.41-261

Uanhoro, J. (2025). ssrhom: Hierarchical ordinal models for analyzing single subject designs [Computer software manual]. Retrieved from `https://jamesuanhoro.r-universe.dev/ssrhom` (R package version 0.0.3.9003)

Valle, D., Ben Toh, K., Laporta, G. Z., & Zhao, Q. (2019, February). Ordinal regression models for zero-inflated and/or over-dispersed count data. *Scientific Reports*, *9*(1), 1–12. doi: 10.1038/s41598-019-39377-x

Vargha, A., & Delaney, H. D. (2000, January). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. doi: 10.3102/10769986025002101

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021, June). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, *16*(2), 667–718. doi: 10.1214/20-BA1221

Wedderburn, R. W. M. (1974, December). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, *61*(3), 439–447. doi: 10.1093/biomet/61.3.439

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0* (Tech. Rep.). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Retrieved from `https://ies.ed.gov/ncee/wwc/Handbooks`

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine*, *12*, 2257–2271.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010, May). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi: 10.1177/0022466908328009

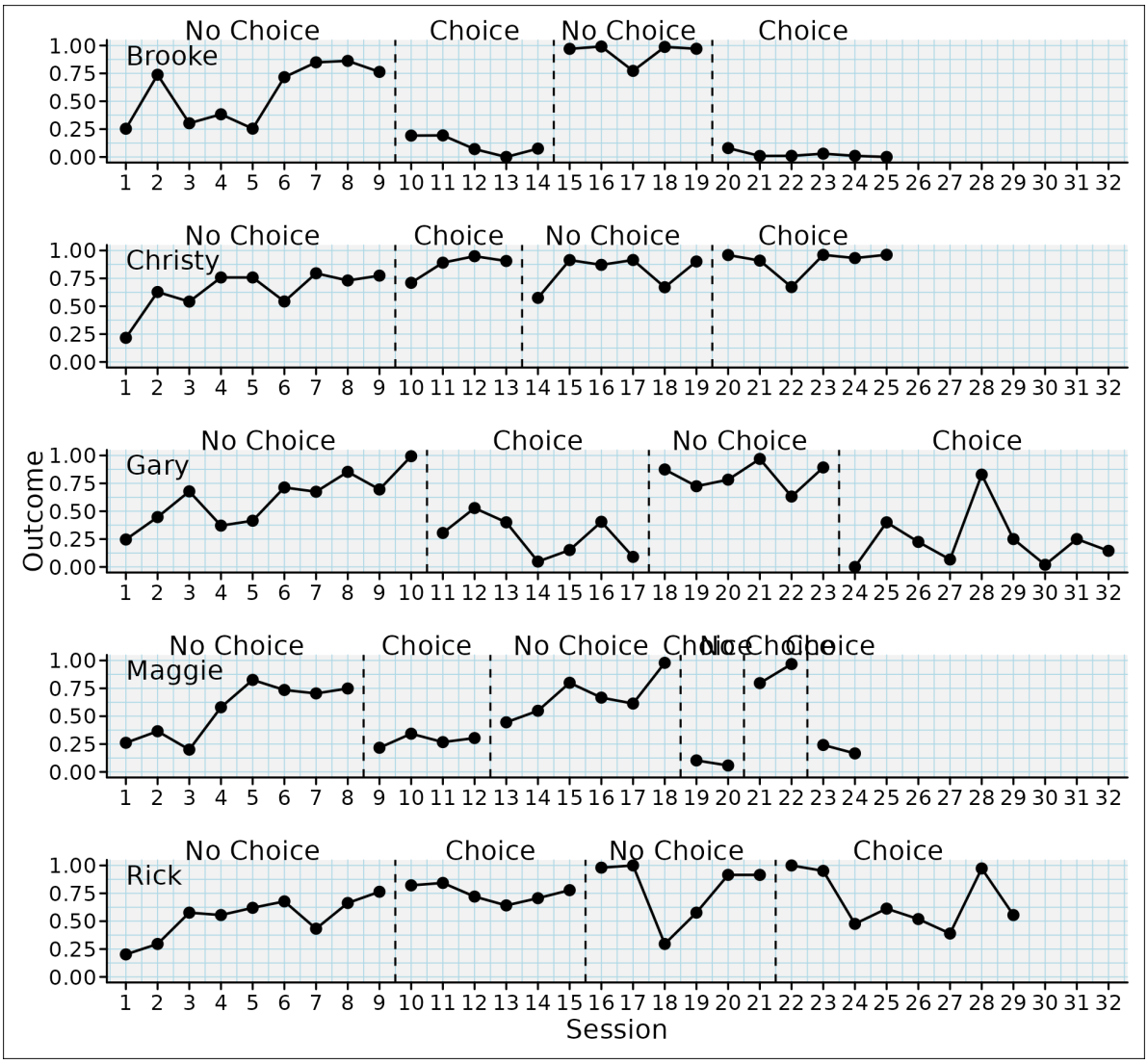
Wooldridge, J. M. (2010). Negative binomial regression models. In *Econometric analysis of cross section and panel data* (pp. 657–660). Cambridge, MA: MIT Press.

Appendix A

Additional results from data demonstration

Figure A1

Romaniuk et al. (2002): Time series for five cases

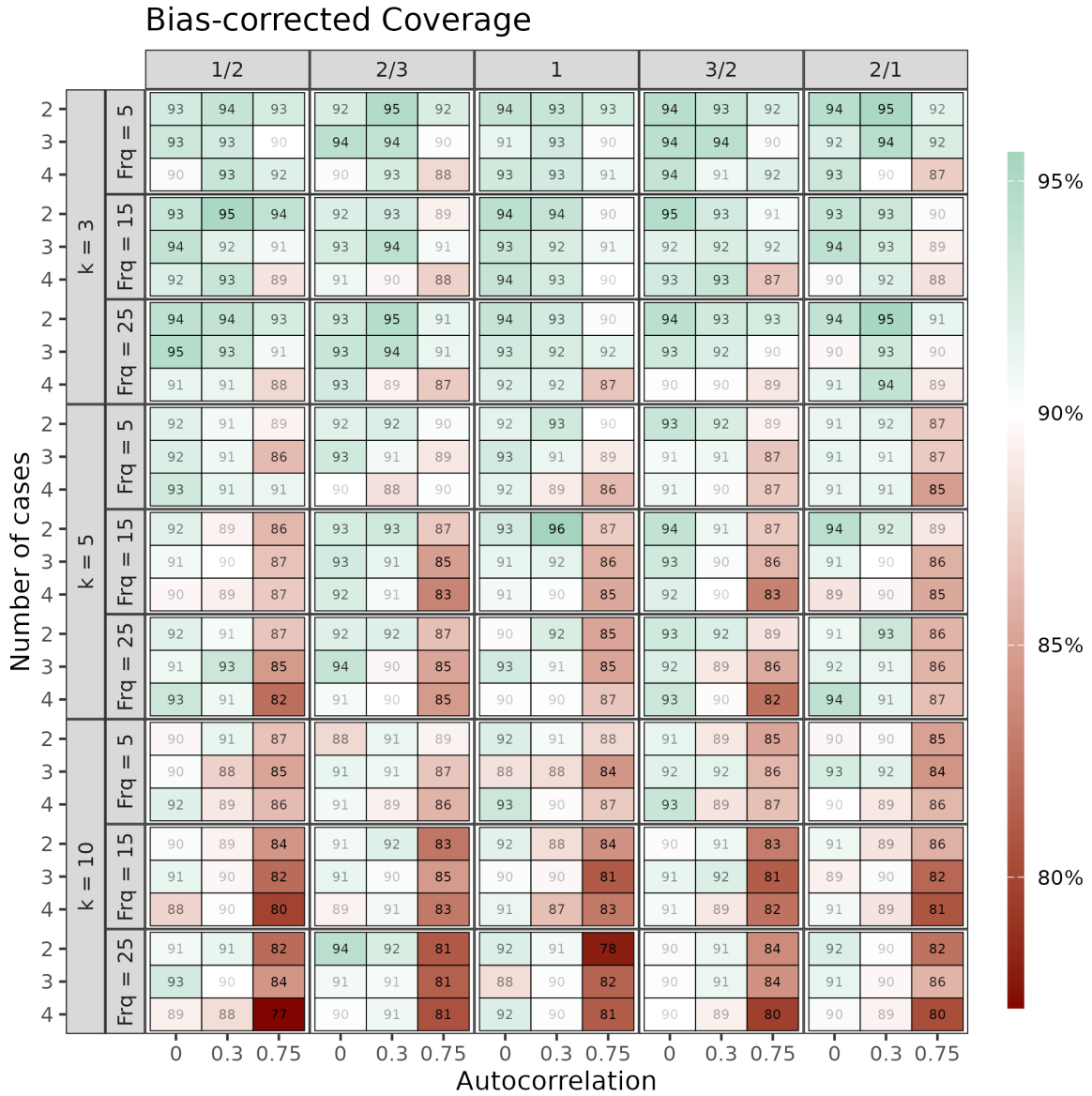


Appendix B

Additional results from Simulation Study 1

Figure B1

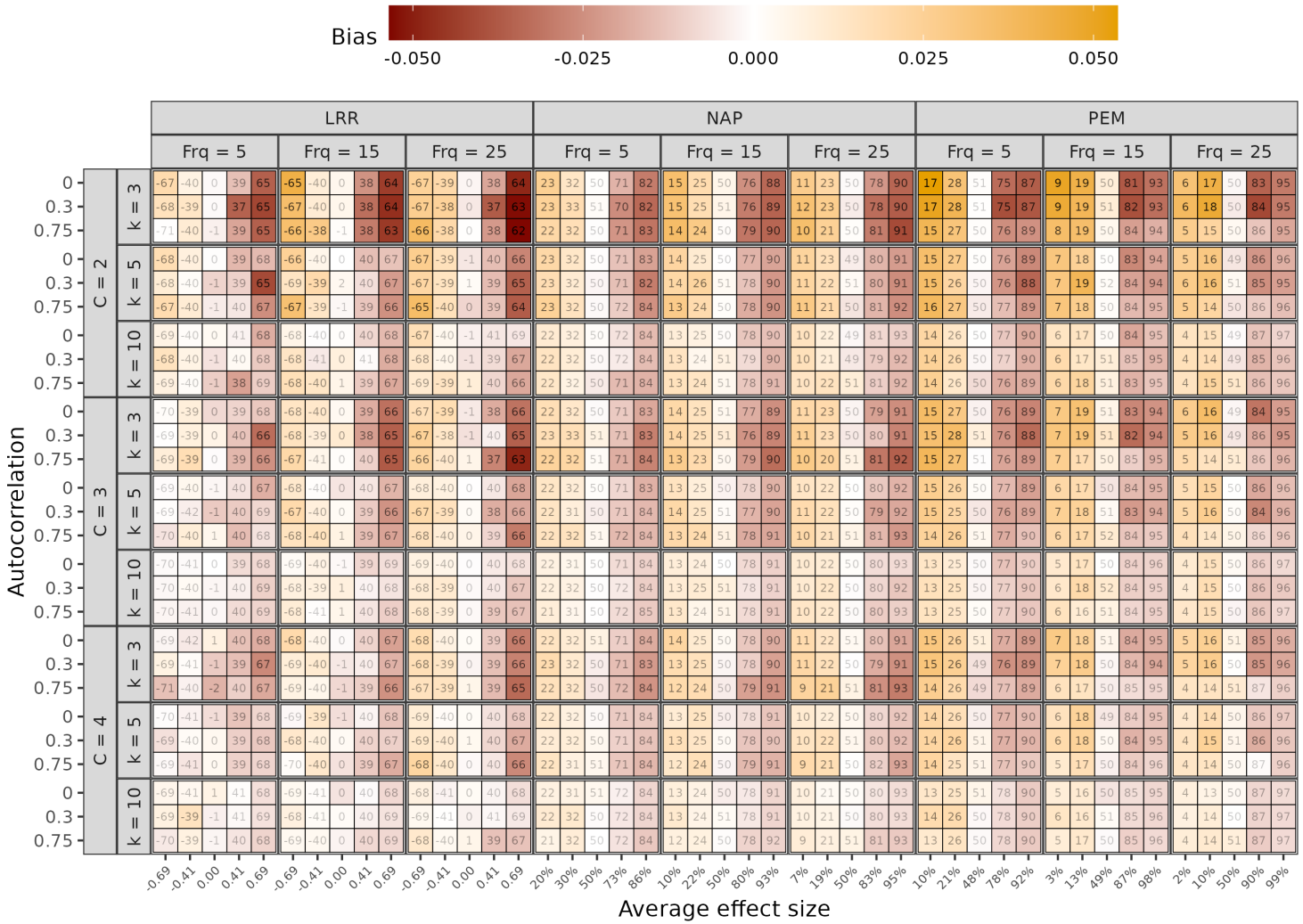
Simulation Study 1: Autocorrelation estimation



Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text.

Figure B2

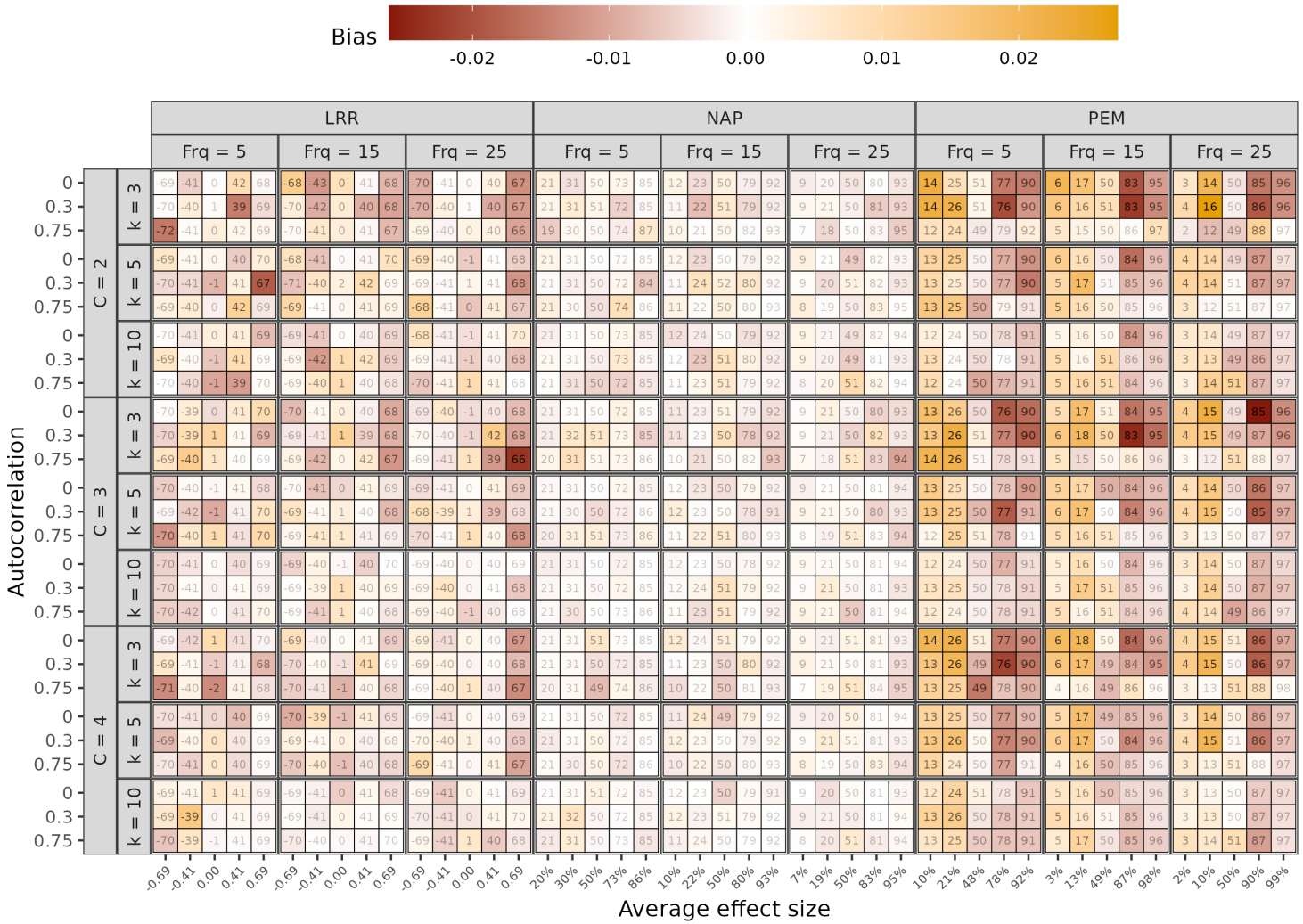
Simulation Study 1: Bias of HOM effect sizes



Note. Average estimate is reported in each cell. Cells with smaller bias have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 6 for population parameters.

Figure B3

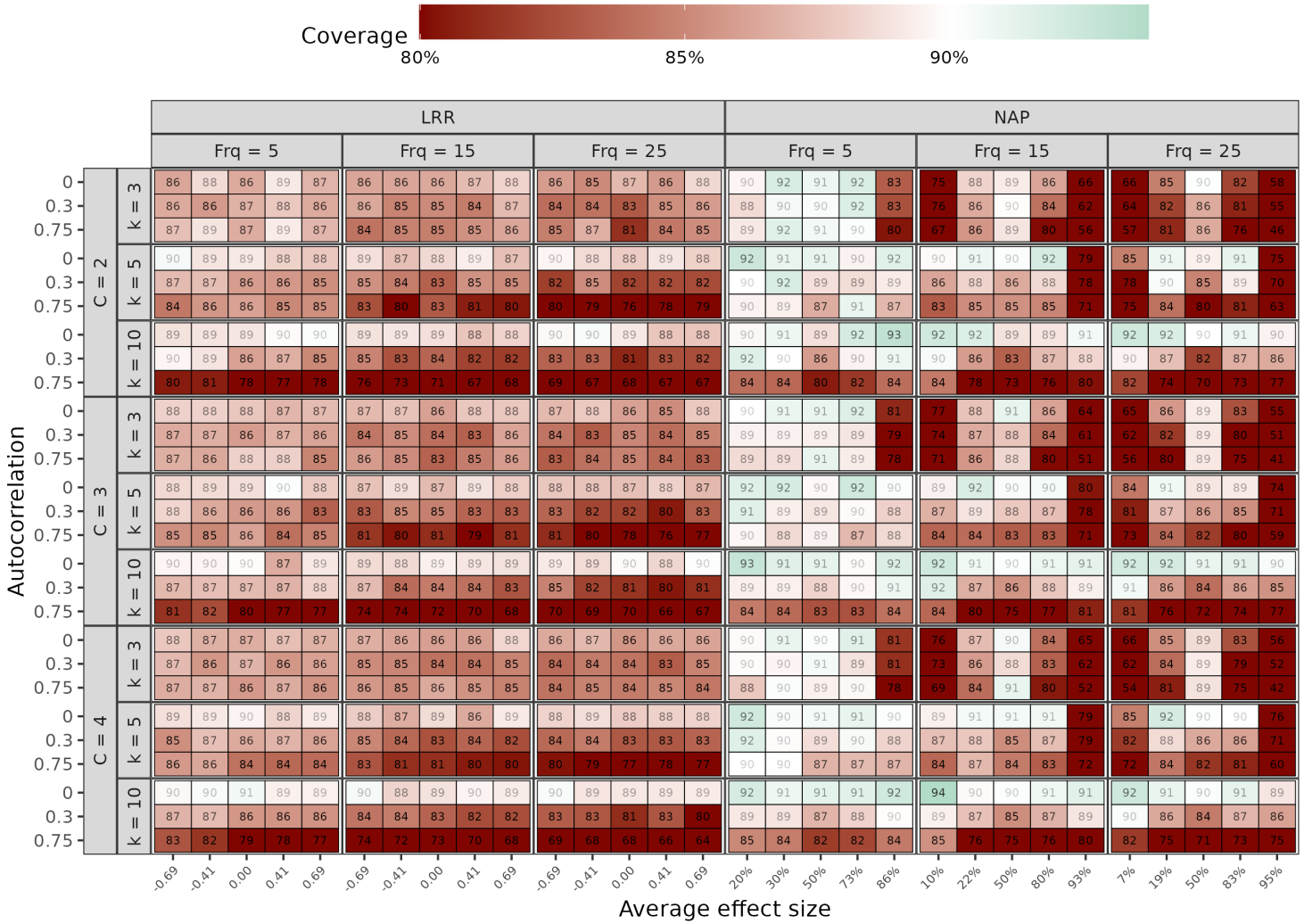
Simulation Study 1: Bias of standard effect sizes



Note. Average estimate is reported in each cell. Cells with smaller bias have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 6 for population parameters.

Figure B4

Simulation Study 1: Empirical coverage of 90% confidence intervals of standard effect sizes



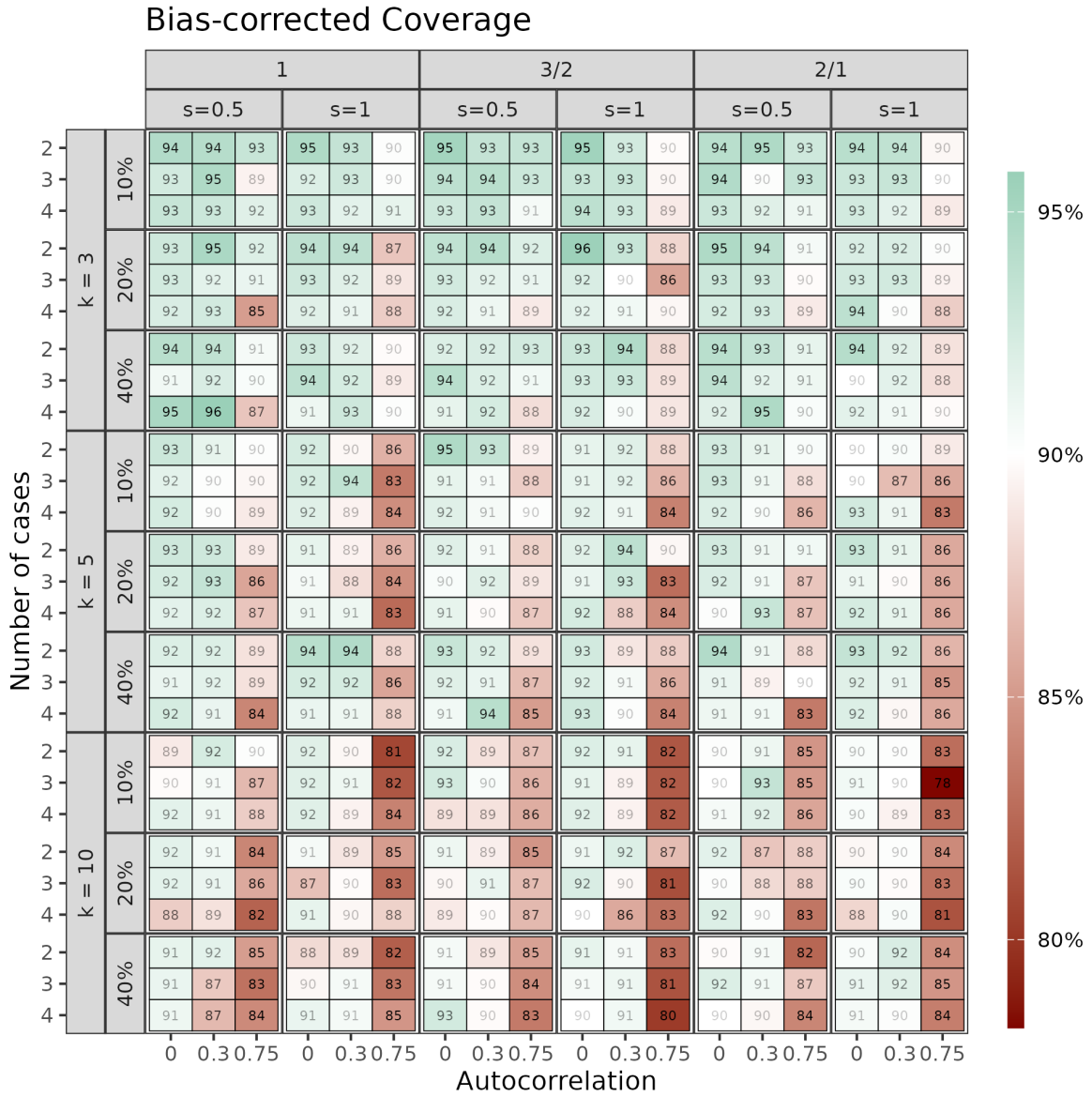
Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 6 for population parameters.

Appendix C

Additional results from Simulation Study 2

Figure C1

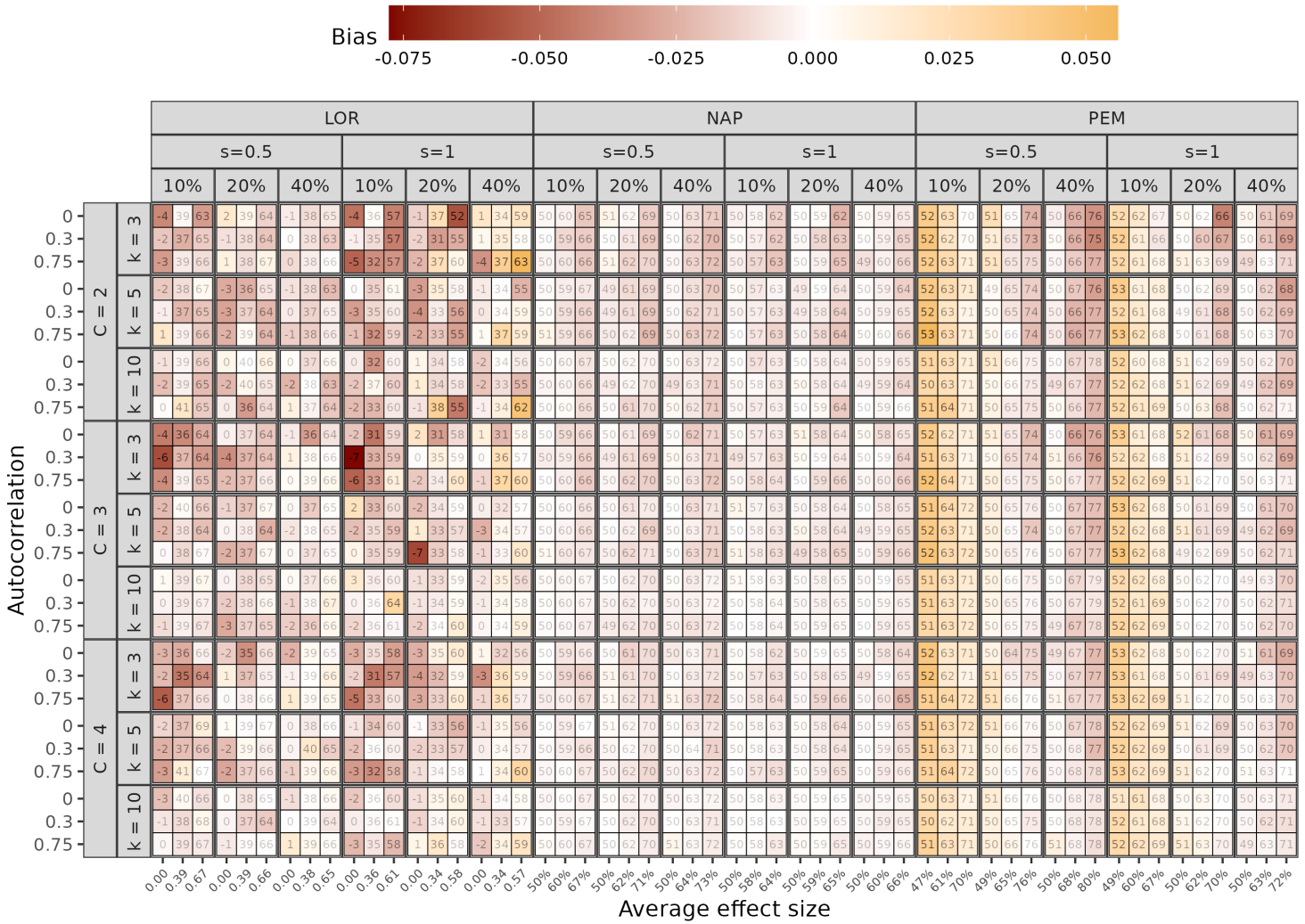
Simulation Study 2: Autocorrelation estimation



Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text.

Figure C2

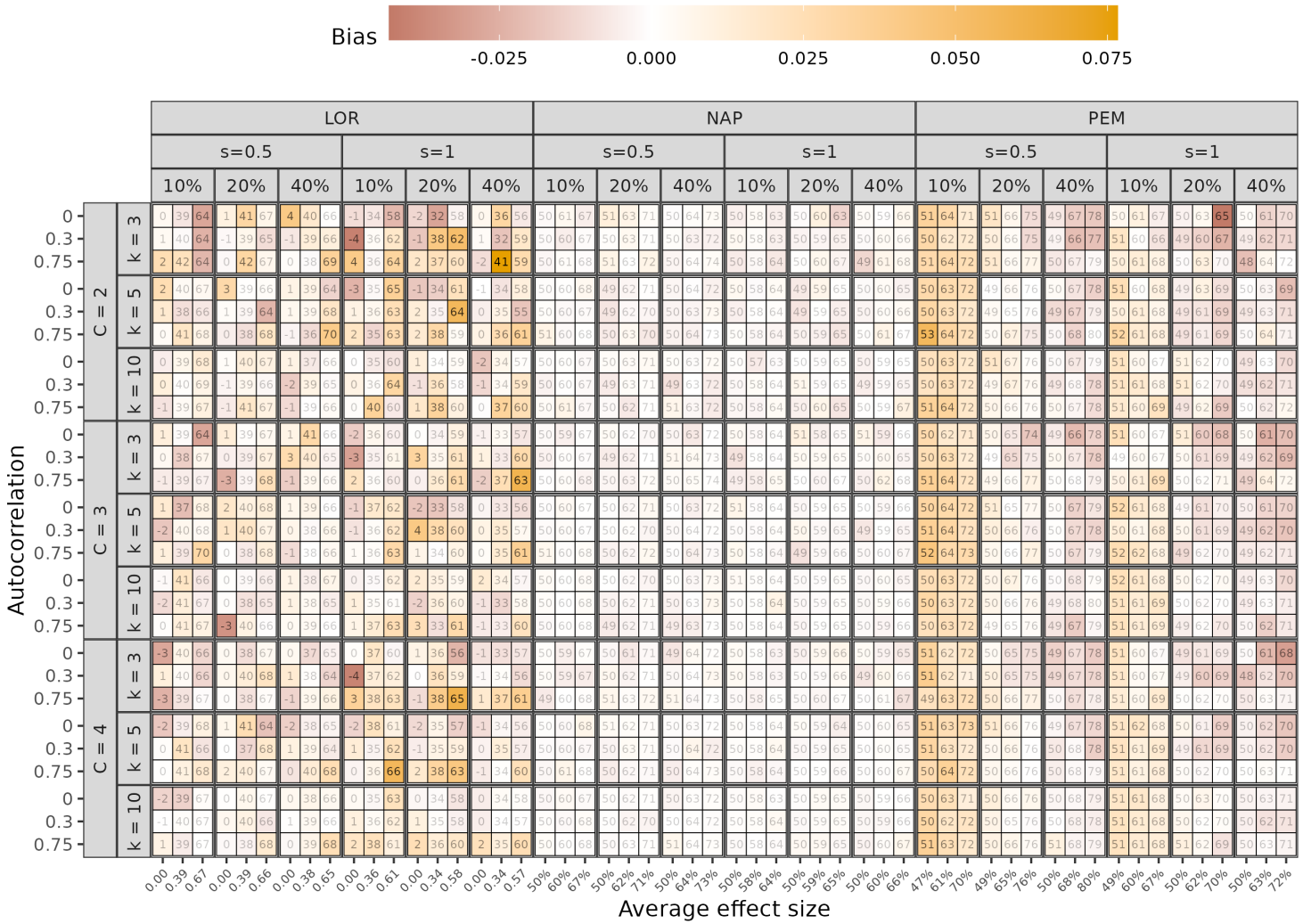
Simulation Study 2: Bias of HOM effect sizes



Note. Average estimate is reported in each cell. Cells with smaller bias have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 9 for population parameters.

Figure C3

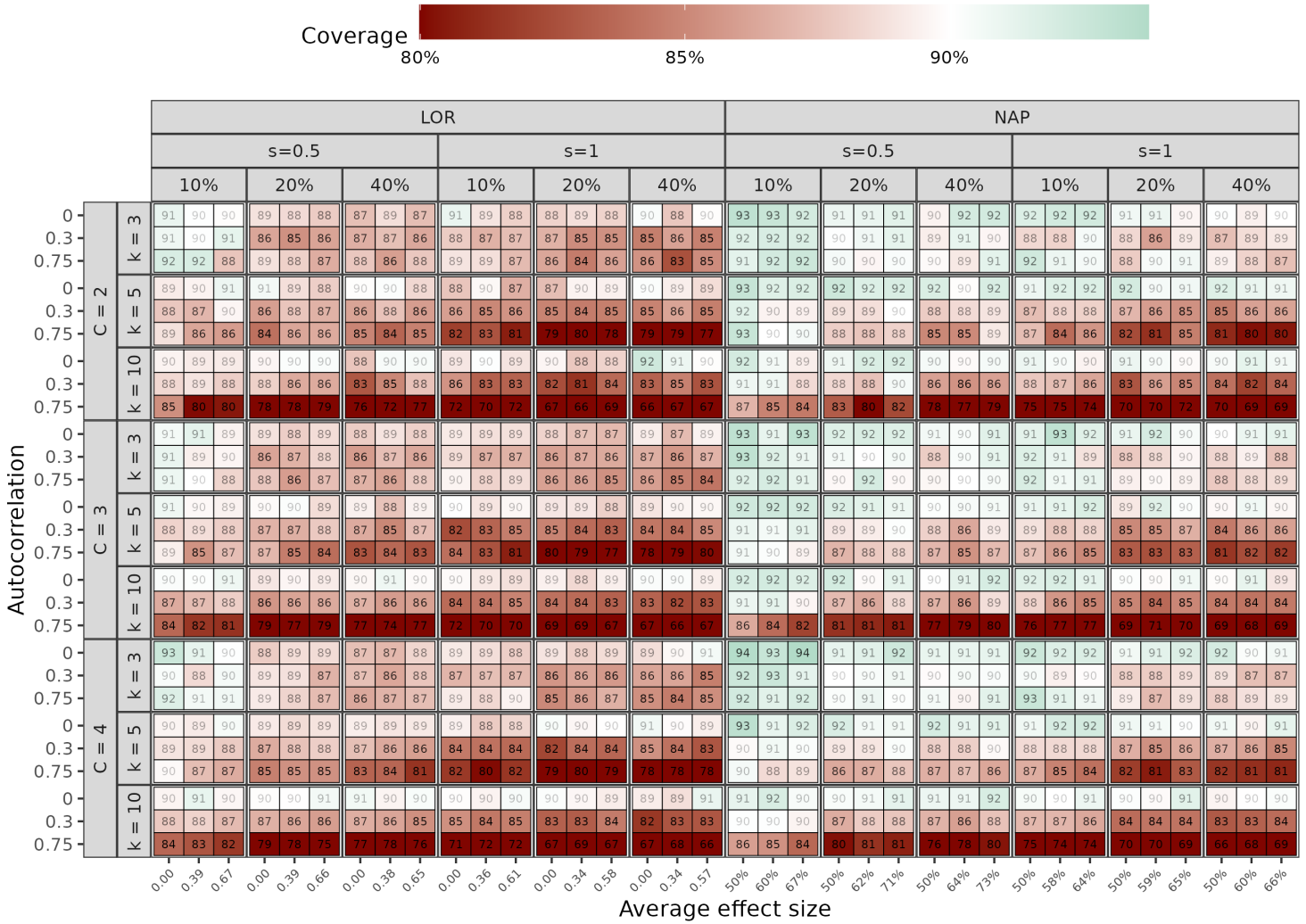
Simulation Study 2: Bias of standard effect sizes



Note. Average estimate is reported in each cell. Cells with smaller bias have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 9 for population parameters.

Figure C4

Simulation Study 2: Empirical coverage of 90% confidence intervals of standard effect sizes



Note. Coverage is reported in each cell. Cells with adequate coverage have fainter text. Note that effect sizes on the x-axis are averaged over some conditions, though the amount of averaging is minor. See Figure 9 for population parameters.