Synthesis of non-overlap of all pairs using logistic transformation or binomial generalized linear mixed models

James E. Pustejovsky[1] & Man Chen[1]

[1] University of Wisconsin-Madison

Author Note

James E. Pustejovsky, Department of Educational Psychology, University of Wisconsin-Madison. Man Chen, Department of Educational Psychology, University of Wisconsin-Madison.

This work was supported, in part, by the Institute of Educational Sciences, U.S. Department of Education through grant R324U190002. The opinions expressed are those of the authors and do not represent the views of the Institute of the U.S. Department of Education.

Correspondence concerning this article should be addressed to James E. Pustejovsky, 1082C Educational Sciences, 1025 W Johnson St. Madison, WI 53706-1706. E-mail: pustejovsky@wisc.edu

Abstract

Available methods for meta-analysis of findings from single-case designs include one-stage methods involving modeling of raw data from across multiple studies and two-stage methods involving calculation of effect sizes and subsequent meta-analysis. The two-stage approach works well for some effect size measures, such as log response ratios, but performs inadequately for the non-overlap of all pairs index. NAP is an effect size in the family of non-overlap measures, which quantify effect magnitude in terms of pairwise rank comparisons of outcomes under different treatment conditions, and is thus a useful metric for outcomes that are not normally distributed and not on a ratio metric. We examine two alternative approaches to meta-analysis of NAP, based on either transforming the effect size estimates or on a binomial generalized linear mixed model. We demonstrate the approaches by re-analyzing data from a meta-analysis of SCEDs examining augmentative and alternative communication interventions and evaluate the performance of the approaches using an extensive simulation study. We find that neither approach performs adequately for synthesis of single-case data series with limited numbers of observations in the baseline and intervention phases.

Keywords: single-case design; non-overlap measure; meta-analysis; generalized linear mixed model

Synthesis of non-overlap of all pairs using logistic transformation or binomial generalized linear mixed models

Meta-Analysis of Single-Case Designs

Methods for meta-analysis of single-case experimental designs (SCEDs) are needed due to the prevalence of empirical research using SCEDs within special education, school psychology, speech and hearing sciences, and other fields. Available meta-analytic approaches include one-stage methods involving modeling of raw data from across multiple studies (Moeyaert et al., 2014; Van den Noortgate & Onghena, 2008) and two-stage methods involving calculation of effect sizes and subsequent meta-analysis (Pustejovsky & Ferron, 2017). The former, raw-data synthesis approach is appropriate when all studies to be synthesized either use the same outcome measure or use outcome measures that can be re-scaled to a common metric. The latter, two-stage approach is appropriate when the studies to be synthesized use a variety of outcome measures, but where an effect size can be estimated for each case within each study. Effect size estimates are then synthesized using a multi-level meta-analytic model that captures variation within and between studies.

A previous simulation study found that the two-stage approach performed well for some effect size measures but not for others (Chen & Pustejovsky, 2021). In particular, a multi-level meta-analysis model with robust variance estimation worked well for log-response ratio effect sizes (Pustejovsky, 2018, 2015) but not for within-case standardized mean differences (Gingerich, 1984) or non-overlap of all pairs (NAP, Parker & Vannest, 2009) effect size measures. Thus, there is an outstanding need for meta-analytic methods that account for the specific properties of the effect size measure.

NAP is an effect size in the family of non-overlap measures, which quantify effect magnitude in terms of pairwise rank comparisons of outcomes under different treatment conditions. Because it is based on rank (ordinal) comparisons, NAP is a useful metric for outcomes that are not normally distributed and not on a ratio metric. The scale of NAP ranges from 0 to 1, with a value of 0.5 corresponding to no effect. The limited range of NAP, along with the strong association between its magnitude and sampling variance, presents challenges for multi-level meta-analysis with normally distributed random effects.

In this study, we study two alternative approaches to meta-analysis of NAP, based on either transforming the effect size estimates or on a binomial generalized linear mixed model. The approach based on transformation of effect sizes is conventional for other effect size measures, but may suffer from problems when the effect size estimates are based on small samples of data. Ryu & Agresti (2008) proposed an approach for combining and comparing NAP values using a binomial generalized linear model with logistic link. We extend their approach to account for the hierarchical structure of NAP effect size estimates by including random effects for each study and for each case. Effect size estimates are modeled as approximately binomially distributed, conditional on the effect size parameter, with a weight function approximated by the variance estimator proposed by Hanley & McNeil (1982). We demonstrate both approaches by re-analyzing data from a meta-analysis of SCEDs examining augmentative and alternative communication interventions for participants with autism spectrum disorders (Ganz et al., 2021). We then evaluate the performance of the approaches using an extensive simulation study, which focuses on the bias and accuracy of the overall average intervention effect estimator and the variance component estimators at each level of the model.

The Sampling Distribution of NAP

Non-overlap of all pairs is one of a number of non-overlap indices that have have been proposed for describing effect size magnitude in the context of single-case research designs (Parker et al., 2014). Like other non-overlap indices, NAP is defined in terms of ordinal comparisons between pairs of outcomes in different conditions. For comparing a baseline condition to an intervention condition, it is the proportion of all possible pairs of observations from the two phases where the outcome from the intervention constitutes a therapeutic improvement over the outcome in baseline (Parker & Vannest, 2009). Parker & Vannest (2009) argued that NAP has better properties than other non-overlap measures because it uses all possible pairs of outcomes, and is therefore more stable and less influenced by outliers than other indices.

Under the assumption that outcomes within a given condition have constant means and standard deviations, NAP is an estimator of a stable parameter measuring the overlap between the distribution of outcomes in the treatment phase, Y^B, and the distribution of outcomes in the baseline phase, Y^A. The definition of the parameter depends on whether therapeutic benefit corresponds to an increase or decrease in the outcome. For an outcome where increase is beneficial, the NAP parameter is

$$\theta = \Pr\left(Y^B > Y^A\right) + \frac{1}{2}\Pr\left(Y^B = Y^A\right), \tag{1}$$

whereas for an outcome where decrease is desirable,

$$\theta = \Pr\left(Y^B < Y^A\right) + \frac{1}{2}\Pr\left(Y^B = Y^A\right) \tag{2}$$

(Pustejovsky, 2019).

Let n_A and n_B denote the number of observations in the baseline and intervention phases, respectively. Let y_s^A denote the s^{th} value of the outcome in the baseline condition, for $s = 1, ..., n_A$, and let y_t^B denote the t^{th} value of the outcome in the intervention condition, for $t = 1, ..., n_B$. The sample estimator of NAP can be defined based on a set of overlap indicator variables, denoted as q_{st} for $s = 1, ..., n_A$ and $t = 1, ..., n_B$, where $q_{st} = 1$ if y_t^B is an improvement over y_s^A, $q_{st} = \frac{1}{2}$ if $y_t^B = y_s^A$, and $q_{st} = 0$ if y_t^B is a worse outcome than y_s^A. The NAP estimator is then

$$\hat{\theta} = \frac{1}{n_A n_B}\sum_{s=1}^{n_A}\sum_{t=1}^{n_B} q_{st} \tag{3}$$

(Parker & Vannest, 2009; Pustejovsky, 2019). Indices equivalent to NAP have long been used in other areas of application, such as clinical medicine (Acion et al., 2006; Vargha & Delaney, 2000), and methodological research from these areas provides some results that are relevant to NAP.

Just as a binomial random variable is a sum of binary variables with a constant probability of success, the NAP estimator is a mean of identically distributed—but correlated—indicator variables, $q_{11}, ..., q_{n_A n_B}$. The sampling variance of the NAP estimator is

$$\text{Var}\left(\hat{\theta}\right) = \frac{\theta(1-\theta)}{n_A n_B}\left[1 + (n_B - 1)\rho_1 + (n_A - 1)\rho_2\right], \tag{4}$$

where

$$\rho_1 = \frac{\text{Cov}(q_{st}, q_{s't})}{\theta(1-\theta)}, \quad \text{and} \quad \rho_2 = \frac{\text{Cov}(q_{st}, q_{st'})}{\theta(1-\theta)}$$

for $s \neq s'$ and $t \neq t'$ (Mee, 1990). Because the indicator variables are not mutually independent and $\rho_1 \geq 0, \rho_2 \geq 0$, the sampling variance of NAP will generally exceed the variance of a binomial distribution.

Sen (1967; see also Mee, 1990) derived an unbiased estimator of the sampling variance of NAP, which can be calculated as

$$V^{Sen} = \frac{1}{(n_A - 1)(n_B - 1)}\left[\hat{\theta}(1-\hat{\theta}) + n_B Q_1 + n_A Q_2 - 2Q_3\right], \tag{5}$$

where

$$\begin{aligned}
Q_1 &= \frac{1}{n_A n_B^2}\sum_{s=1}^{n_A}\left[\sum_{t=1}^{n_B}\left(q_{st} - \hat{\theta}\right)\right]^2 \\
Q_2 &= \frac{1}{n_A^2 n_B}\sum_{t=1}^{n_B}\left[\sum_{s=1}^{n_A}\left(q_{st} - \hat{\theta}\right)\right]^2 \\
Q_3 &= \frac{1}{n_A n_B}\sum_{s=1}^{n_A}\sum_{t=1}^{n_B}\left(q_{st} - \hat{\theta}\right)^2.
\end{aligned}$$

This variance estimator has the property that it is equal to zero if $\hat{\theta}$ is equal to zero or one (i.e., when there is no overlap between observations from different conditions). A strictly positive variance estimator can be calculated by replacing $\hat{\theta}$ with the quantity

$$\tilde{\theta} = \begin{cases} \frac{1}{2n_A n_B} & \text{if} \quad \hat{\theta} = 0 \\ \hat{\theta} & \text{if} \quad 0 < \hat{\theta} < 1 \\ \frac{2n_A n_B - 1}{2n_A n_B} & \text{if} \quad \hat{\theta} = 1 \end{cases}$$

in Equation (5). Hanley & McNeil (1982) proposed a slightly simpler variance estimator, given by

$$V^{HM} = \frac{1}{n_A n_B}\left[\tilde{\theta}(1-\tilde{\theta}) + (n_B - 1)Q_1 + (n_A - 1)Q_2\right], \tag{6}$$

where we have again used $\tilde{\theta}$ in place of $\hat{\theta}$ so that V^{HM} is strictly greater than zero. Note that V^{HM} is always smaller than V^{Sen}.

Meta-analysis of NAP

The NAP parameter is bounded between zero and one (inclusive), and the distribution of the NAP estimator $\hat{\theta}$ is far from Gaussian and can be quite skewed, particularly when θ is near the extreme. Furthermore, single-case data typically follow a hierarchical structure, where we have a summary effect size for each case, with several cases nested within each study. All of these features pose challenges for meta-analysis of NAP. A potential solution is to consider meta-analysis of a transformation of the NAP parameter, such as positing a model involving a logistic transformation, which puts the effect size parameter on an unconstrained scale. We also use a hierarchical meta-analysis model to account for the potential dependence arising from cases nested within studies.

Let θ_{jk} denote the NAP parameter for case j from study k, where $j = 1, ..., J_k$ and $k = 1, ..., K$. We consider meta-analytic models of the form

$$\operatorname{logit}(\theta_{jk}) = \mu + u_k + v_{jk}, \tag{7}$$

where μ is the overall average effect size (on the logistic scale) across cases and studies, $u_k \sim N(0, \tau^2)$ is a study-level random effect and $v_{jk} \sim N(0, \omega^2)$ is a case-level random effect. This model describes the distribution of true effect size parameters across cases and studies. To complete the model, we need to make further assumptions describing how the distribution of effect size estimates relates to the case-specific parameters. We consider two approaches to doing so, one based on transforming the NAP estimates and one based on a generalized linear mixed model.

Transforming the NAP estimator

In other realms of research synthesis, it is common to conduct meta-analysis after making a transformation of effect size estimates, such as using Fisher's z-transformation of the Pearson correlation coefficient. One could apply the same approach here by taking the logistic transformation of the NAP estimator. However, this transformation is undefined at the extremes of the scale, and so we use the truncated version of NAP so that the transformed effect size estimator remains well-defined. Thus, we can meta-analyze $\text{logit}(\tilde\theta_{jk})$. Following the usual delta method, the variance of this estimator is approximately

$$\text{Var}\left(\text{logit}(\tilde\theta)\right) \approx \frac{1}{\theta^2(1-\theta)^2} \times \text{Var}(\hat\theta),$$

which we estimate by substituting $\tilde\theta$ for θ and V^{HM} or V^{Sen} for $\text{Var}(\hat\theta)$. The full meta-analysis model is then

$$\text{logit}(\tilde\theta_{jk}) = \mu + u_k + v_{jk} + e_{jk}, \tag{8}$$

where we assume that e_{jk} has mean zero and variance $V_{jk} / \left[\tilde\theta_{jk}^2(1-\tilde\theta_{jk})^2\right]$, which is treated as a known quantity.

We would expect this approach to work reasonably well if each NAP estimate is based on a comparison of phases with a large number of observations. However, single-case designs often include a very limited number of observations in each phase. As a result, the transformed effect size estimator might have a non-trivial bias and the delta-method variance approximation might be inadequate. Furthermore, it might not be reasonable to treat the sampling variances as known quantities. We assess the extent of these problems in the simulation study.

Generalized linear mixed model

Mee (1990) considered approximating the sampling distribution of the NAP estimator as

$$n_A n_B \hat{\theta} \overset{\cdot}{\sim} Binom(\theta, \tilde{N}), \tag{9}$$

where $\tilde{N}$ is the ratio of $\theta(1-\theta)$ to an estimate of the variance of the NAP estimator. Using the Hanley & McNeil (1982) variance estimator, this gives

$$\tilde{N}^{HM} = \frac{n_A n_B}{1 + (n_B - 1)\frac{Q_1}{\tilde{\theta}(1-\tilde{\theta})} + (n_A - 1)\frac{Q_2}{\tilde{\theta}(1-\tilde{\theta})}}.$$

This quantity can be interpreted as the “effective” number of trials for the NAP estimator, that is, the number of trials needed for a binomial distribution to have variance equal to the variance of NAP.

Now, let n_{Ajk}, n_{Bjk}, $\tilde{N}_{jk}^{HM}$, and $\hat{\theta}_{jk}$ denote the sample sizes, effective number of trials, and unbiased NAP estimator for case j from study k, for $j = 1, ..., J_k$ and $k = 1, ..., K$. A binomial generalized linear mixed model for NAP can be described in two parts. First, at the measurement level, we assume

$$n_{Ajk} n_{Bjk} \hat{\theta}_{jk} \sim Binom\left(\theta_{jk}, \tilde{N}_{jk}^{HM}\right). \tag{10}$$

Second, we assume that the NAP parameters follow a multi-level model with random effects at the case level and study level, all on the logistic scale, as given in Equation (7).

Combining Equations (10) and (7) yields a binomial generalized linear mixed model with logistic link function. We estimate the model using approximate maximum likelihood via Laplace approximation, as implemented in the glmmTMB package (Brooks et al., 2017).

The generalized linear mixed model formulation uses a binomial likelihood to describe the distribution of $\hat{\theta}$ conditional on the true parameter. This has the advantages of capturing the skew of the sampling distribution and allowing us to avoid truncation of the estimator at the extremes of the distribution. However, this approach also bears the potential limitation that the effective number of trials is treated as known, when in fact it must be *estimated* for each case. In practice, the effective number of trials is estimated based on a small sample of data for each case, which might create problems for this approach to synthesis. Thus, we need to investigate its performance using Monte Carlo simulation.

Prediction intervals

Prediction intervals are a useful technique for characterizing the degree of heterogeneity in meta-analysis results (Borenstein et al., 2017; Brannick et al., 2021). A $B \times 100\%$ prediction interval is an interval estimate that is expected to contain a *new* effect size with probability B, or equivalently, that is expected to contain $B \times 100\%$ of the overall distribution of effect sizes, on average. In multi-level meta-analysis of single-case designs, we can distinguish between study-level prediction intervals and case-level prediction intervals.

The study-level prediction interval is constructed with respect to the distribution of study-level average effect sizes, so that it will contain a new study-level average effect size with probability B. We calculate an approximate $B \times 100\%$ study-level prediction interval as

$$\hat{\mu} \ \pm z_{B/2} \times \sqrt{SE_{\hat{\mu}}^2 + \hat{\tau}^2}, \tag{11}$$

where $\hat{\mu}$ is the estimate of the overall average effect size (on the logistic scale), $SE_{\hat{\mu}}$ is its standard error, $\hat{\tau}$ is the estimated between-study standard deviation, and $z_{B/2}$ is the upper $B/2$ critical value from a standard normal distribution.

The case-level prediction interval is constructed with respect to the distribution of cases, so that it will contain the effect size for a new case from a new study with probability B. We calculate an approximate $B \times 100\%$ case-level prediction interval as

$$\hat{\mu} \ \pm z_{B/2} \times \sqrt{SE_{\hat{\mu}}^2 + \hat{\tau}^2 + \hat{\omega}^2}, \tag{12}$$

where $\hat{\omega}$ is the estimated within-study standard deviation.

One advantageous feature of prediction intervals is that they can be reported either on the logistic scale (i.e., the scale on which the meta-analysis model is defined) or on the original NAP scale. The intervals defined above are on the logistic scale. To put the interval on the original scale, we apply the inverse of the logistic transformation to the end-points of the interval.

Meta-analysis of augmentative and alternative communication interventions

Ganz et al. (2021) reported a synthesis of single-case design studies examining the effects of augmentative and alternative communication (AAC) interventions for school-age individuals with autism spectrum disorders (ASD) or intellectual disabilities. The analysis by Ganz and colleagues used two different effect size metrics, Tau (a linear transformation of NAP) and the log response ratio, and synthesized findings using multi-level meta-analytic models. For the analysis of Tau effect sizes, the meta-analytic model was specified on the original scale despite a very non-normal distribution of effect size estimates, and robust variance estimation methods were used for inference. Here, we re-analyze the data using the two approaches described in the previous section.

Ganz et al. (2021) identified over 100 single-case design studies meeting inclusion criteria, including over 300 unique participants. For illustrative purposes, we limited our

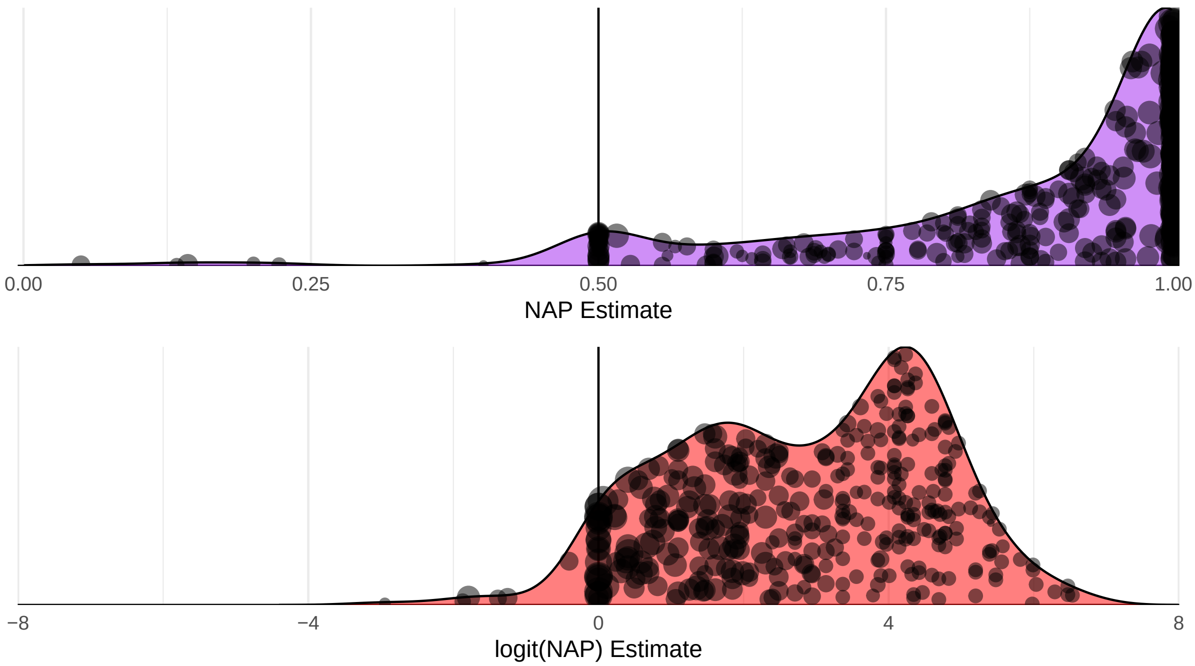

Figure 1

Density of NAP estimates (top panel, purple) and logit(NAP) estimates (bottom panel, red) from AAC intervention studies. Each point corresponds to one NAP estimate, with point size inversely proportional to the log of its sampling variance.

re-analysis to participants with ASD and outcome measures directly related to AAC. With these additional criteria, the sample included data for 172 distinct participants from 65 studies. Included studies used several different designs, including multiple baseline across participants, treatment reversal, and alternating treatment designs. For treatment reversal designs, we calculated NAP effect sizes for each pair of adjacent baseline and treatment phases (i.e., each consecutive AB pair) and treat the resulting effect size estimates as independent replicates of the same parameter. For alternating treatment designs involving multiple interventions or comparison conditions, we calculated NAP effect sizes comparing conditions with AAC interventions to baseline conditions. The data include a total of 366 effect size estimates.

Figure 1 displays the distribution of NAP effect size estimates, with the original (0,1) scale depicted in the top panel and the logistic scale depicted in the bottom panel.

The size of each point is inversely proportional to the log of the effect size's sampling variance, such that larger points correspond to more precise estimates. In the top panel, it can be seen that the distribution of NAP estimates is strongly left-skewed and the truncation of effect sizes near the ceiling of 1 is evident.

Table 1
Multi-level meta-analysis estimates for augmentative and alternative communication intervention studies

	Logistic-transformation		Binomial GLMM	
Parameter	Est	95% CI	Est	95% CI
mu	2.213	[1.858, 2.569]	2.945	[2.394, 3.496]
NAP	0.901	[0.865, 0.929]	0.950	[0.916, 0.971]
tau	1.305	[1.039, 1.641]	1.051	[0.834, 1.324]
omega	0.590	[0.413, 0.790]	2.032	[1.590, 2.596]

Table 1 reports parameter value estimates for the multi-level meta-analysis model using both the logistic transformation approach and the binomial GLMM approach. With the former approach, the average effect size estimate of $\hat{\mu} = 2.213$ corresponds to a value of 0.901 on the original NAP scale. The binomial GLMM approach leads to a larger average effect size estimate of $\hat{\mu} = 2.945$, corresponding to a value of 0.950 on the original NAP scale. The estimated study-level standard deviation is smaller when based on the binomial GLMM than when based on the logistic transformation. Notably, the estimated case-level standard deviation is substantially larger when based on the binomial GLMM versus when based on the logistic transformation. Figure 2 depicts the distibutions of effect size parameters implied by each estimation approach, along with the empirical distributions of effect size estimates.

The parameter estimates from these two approaches imply substantively different prediction intervals. With the logistic transformation approach, an 80% study-level prediction interval is given by [0.525, 3.901] ([0.628, 0.980] on the NAP scale) and an 80% case-level prediction is [0.364, 4.063] ([0.590, 0.983] on the NAP scale). In contrast, with

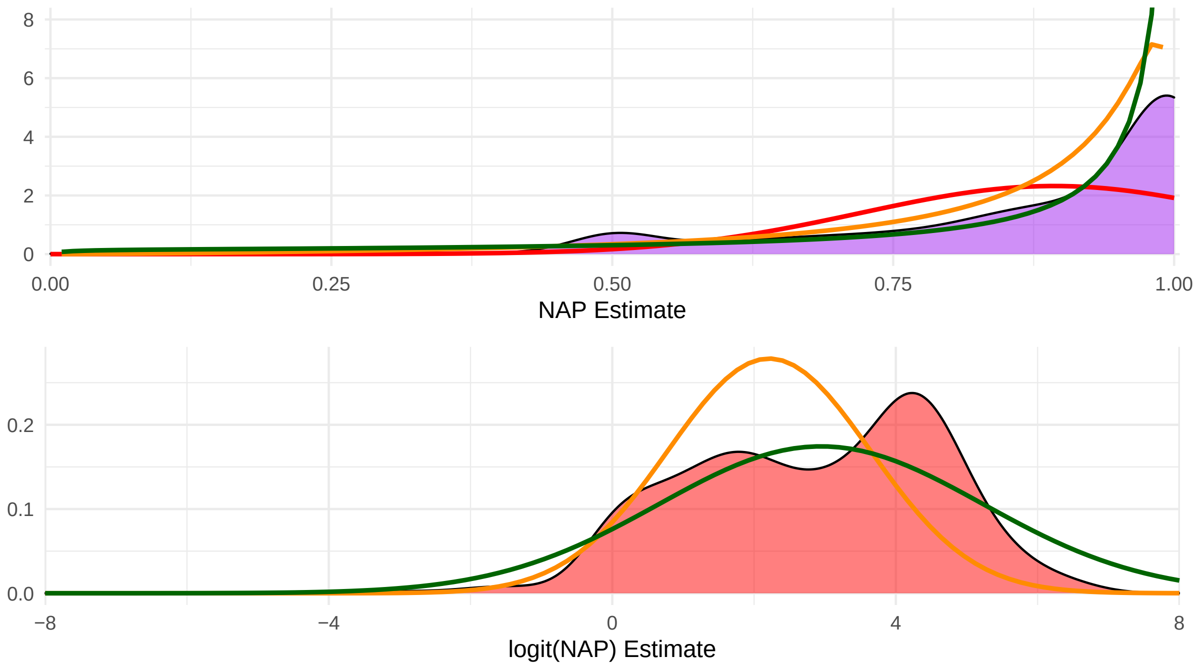

Figure 2

Density of NAP estimates (top panel, purple) and logit(NAP) estimates (bottom panel, red) from AAC intervention studies. Orange curves represent the model-based density estimates based on the logistic transformation approach. Green curves represent the model-based density estimates based on the binomial GLMM approach. The red curve in the top panel represents the model-based density estimate based on a multi-level meta-analysis of the raw NAP estimates.

the binomial GLMM approach, the study-level prediction interval is [0.316, 5.574] ([0.578, 0.996] on the NAP scale) and the case-level prediction interval is [-0.009, 5.899] ([0.498, 0.997] on the NAP scale). The wider range of the prediction intervals from the binomial GLMM implies greater uncertainty about the effect sizes that would be expected when using an AAC intervention in a new study.

A multi-level meta-analysis of the raw NAP estimates would seem to be clearly inappropriate given the strongly skewed, non-normal distribution of effect size estimates. For purposes of comparison, we nonetheless computed prediction intervals based on such a model, truncating the intervals at the limits of the NAP scale. The resulting estimates are depicted as a red curve in the top panel of Figure 2. The 80% study-level prediction interval is [0.685, 1.000] and the 80% case-level prediction is [0.672, 1.000]. These intervals differ in both their center and width compared to the prediction intervals obtained from the logistic transformation and binomial GLMM approaches.

The example presented here is based on real data where no ground truth is known. Further, it is difficult to say on the basis of theory which of the two estimation approaches we have used provides more accurate results—or whether either of the approach provides estimates of the distribution of effects that adequately represent uncertainty. To investigate the performance characteristics of the approaches, we turned to Monte Carlo simulations.

Monte Carlo Simulations

Data-generating process

We simulated meta-analytic data using an approach similar to that used in Chen & Pustejovsky (2021), following a strategy of generating raw data from a collection of multiple single-case design studies, then calculating a NAP effect size estimate for each case within each study. We generated data for $K = 10$, 20, or 30 primary studies. Study k included J_k cases, where J_k was drawn from a uniform distribution on the integers 1,. . . ,5. We generated true effect sizes for each case in each study according to Equation (7), with values of the overall average effect size that correspond to NAPs ranging from 0.05 to 0.95 in steps of 0.10; values of τ equal to 0.0, 0.1, 0.2, or 0.3; and values of ω equal to 0.00, 0.05, 0.10, or 0.15.

In order to generate raw data, we also had to specify a model for the distribution of outcomes in the baseline phase for each case, as well as models for the number of observations in the baseline and intervention phases. We assumed that outcomes were either normally (Gaussian) distributed with unit variance or Poisson distributed. In each case, we simulated the baseline mean level of the outcome from a Gamma distribution with shape 2 and scale 7, truncated to have a minimum value of 5, and we assumed that the mean level was the same for all cases within a given study. Letting α_k denote the mean baseline level for cases in study k,

$$\alpha_k \sim \max\left\{5, \Gamma(2,7)\right\}.$$

We truncated this distribution in order to avoid range restrictions for the NAP parameters. We simulated the number of observations in the baseline and intervention phases from shifted Poisson distributions, as

$$n_A \sim 3 + Poisson(4), \qquad n_B \sim 3 + Poisson(4).$$

Given a true effect size for case j in study k and a baseline mean level for study k, the mean level of the outcome in the intervention phase can be determined using the properties of the normal or Poisson distribution (Chen & Pustejovsky, 2021). We then simulated n_{Ajk} raw data points with the specified baseline mean level and n_{Bjk} raw data points with the specified intervention mean level, then calculated a NAP estimate and sampling variance according to equations (3) and (4), respectively.

Estimation methods

For each simulated meta-analytic dataset, we applied the logistic transformation approach and the binomial GLMM approach. With the logistic transformation approach, we estimated the meta-analytic model by restricted maximum likelihood using the metafor package (Viechtbauer, 2010). We estimated the binomial GLMM by restricted maximum likelihood using the glmmTMB package (Brooks et al., 2017) with default settings. Note that, for each approach, estimation involves use of numerical maximization routines that do not always converge.

Performance criteria

For each combination of parameter values, we generated 1000 meta-analytic datasets. We tracked convergence rates and calculated performance criteria for each model based on the subset of replications where the estimation process converged. We assessed the performance of the approaches in terms of the parameter bias of $\hat{\mu}$, $\hat{\tau}^2$, and $\hat{\omega}^2$, the coverage rate and width of 95% confidence intervals for μ, and the expected content of 80% prediction intervals.

We calculated the expected content of prediction intervals based on the true data-generating process. Let $\left[l^r_{study}, u^r_{study}\right]$ and $[l^r_{case}, u^r_{case}]$ denote the lower and upper bounds of the study-level and case-level prediction intervals, respectively, calculated in replication $r = 1, ..., 1000$. The content of the r^{th} study-level prediction interval is

$$C^r_{study} = \Phi\left(\frac{u^r_{study} - \mu}{\tau}\right) - \Phi\left(\frac{l^r_{study} - \mu}{\tau}\right),$$

where $\Phi()$ is the standard normal cumulative distribution function. The content of the r^{th} case-level prediction interval is

$$C^r_{case} = \Phi\left(\frac{u^r_{case} - \mu}{\sqrt{\tau^2 + \omega^2}}\right) - \Phi\left(\frac{l^r_{case} - \mu}{\sqrt{\tau^2 + \omega^2}}\right).$$

We estimated the expected content by taking the average of C^r_{study} or C^r_{case} across replications of the simulation.

Results

For simplicity, we present results for Poisson-distributed outcomes in the figures of the main text. Results for normally distributed outcomes are generally similar and will be presented in supplementary materials. Each of the following figures consists of a grid of panels, where columns correspond to different levels of case-level heterogeneity (ω) and rows correspond to different levels of study-level heterogeneity (τ). Within each panel, the horizontal axis corresponds to different values of the overall average effect size (μ) on the logistic scale, line types correspond to different numbers of studies in the meta-analysis (K), and colors correspond to different estimation approaches.

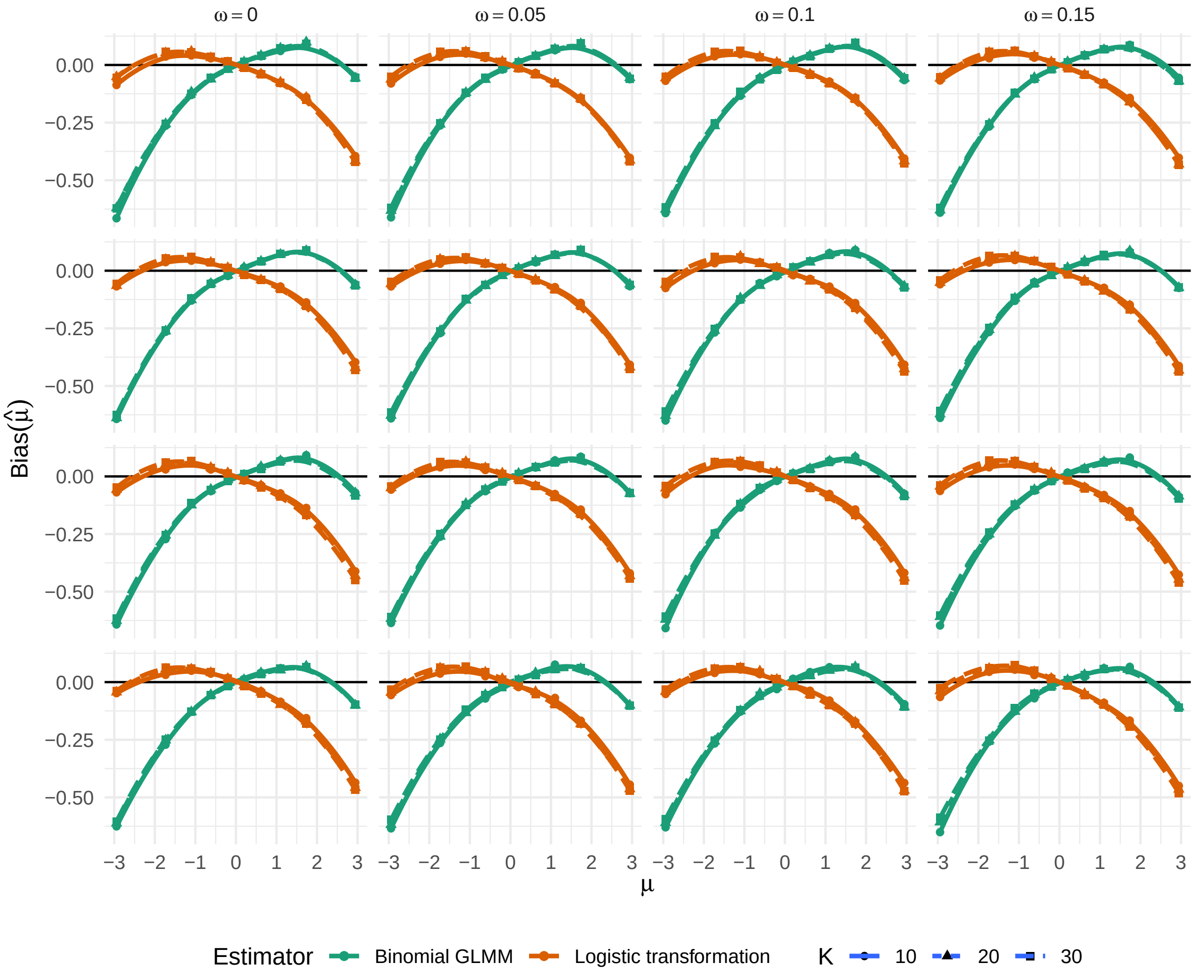

Figure 3

Bias of $\hat{\mu}$ using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

Figure 3 depicts the bias of the overall average effect size estimators ($\hat{\mu}$) based on the logistic transformation (in orange) and binomial GLMM (in green) approaches. Both estimators are systematically biased for non-null μ, with a similar pattern of bias across different levels of within- and between-study heterogeneity. Neither estimator has uniformly smaller bias. Rather, the logistic transformation approach has relatively small bias for negative average effect sizes (i.e., $\mu < 0$ or $NAP < 0.5$) but it is biased towards null for positive average effect sizes, with bias that increases in magnitude for larger average effect sizes. The binomial GLMM estimator shows the opposite pattern of bias: it

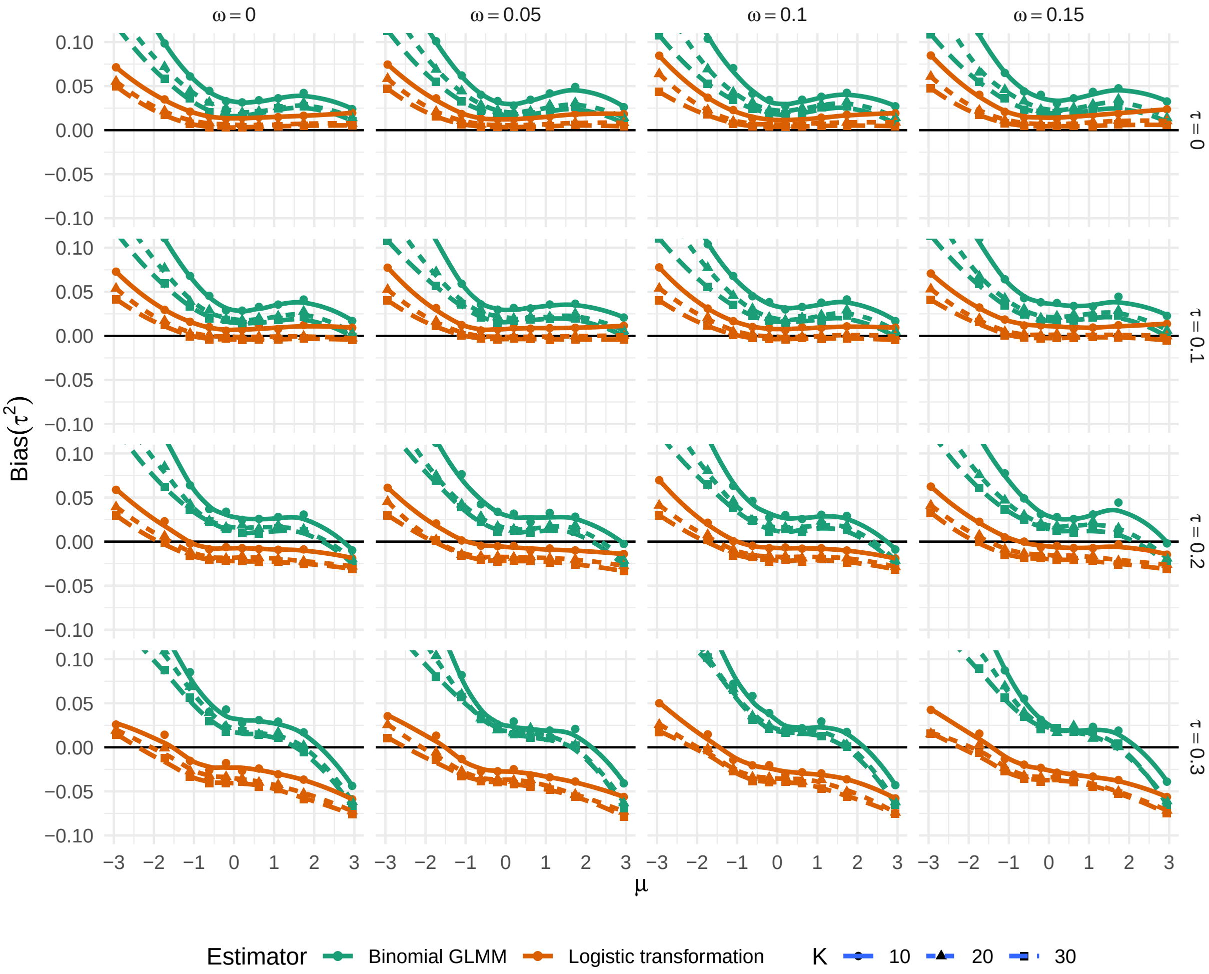

Figure 4

Bias of $\hat{\tau}^2$ using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

has relatively small bias for positive average effect sizes, but it is biased away from zero for negative average effect sizes, and the magnitude of the bias grows as μ becomes increasingly negative.

Figure 4 depicts the bias of the between-study heterogeneity variance estimator ($\hat{\tau}^2$). Overall, the estimator based on logistic transformation tends to be less biased than the estimator based from the binomial GLMM. For smaller values of τ, the logistic transformation has a relatively small, positive bias over most of the parameter space, although its bias becomes negative at the largest value of $\tau = 0.3$. The estimator based on

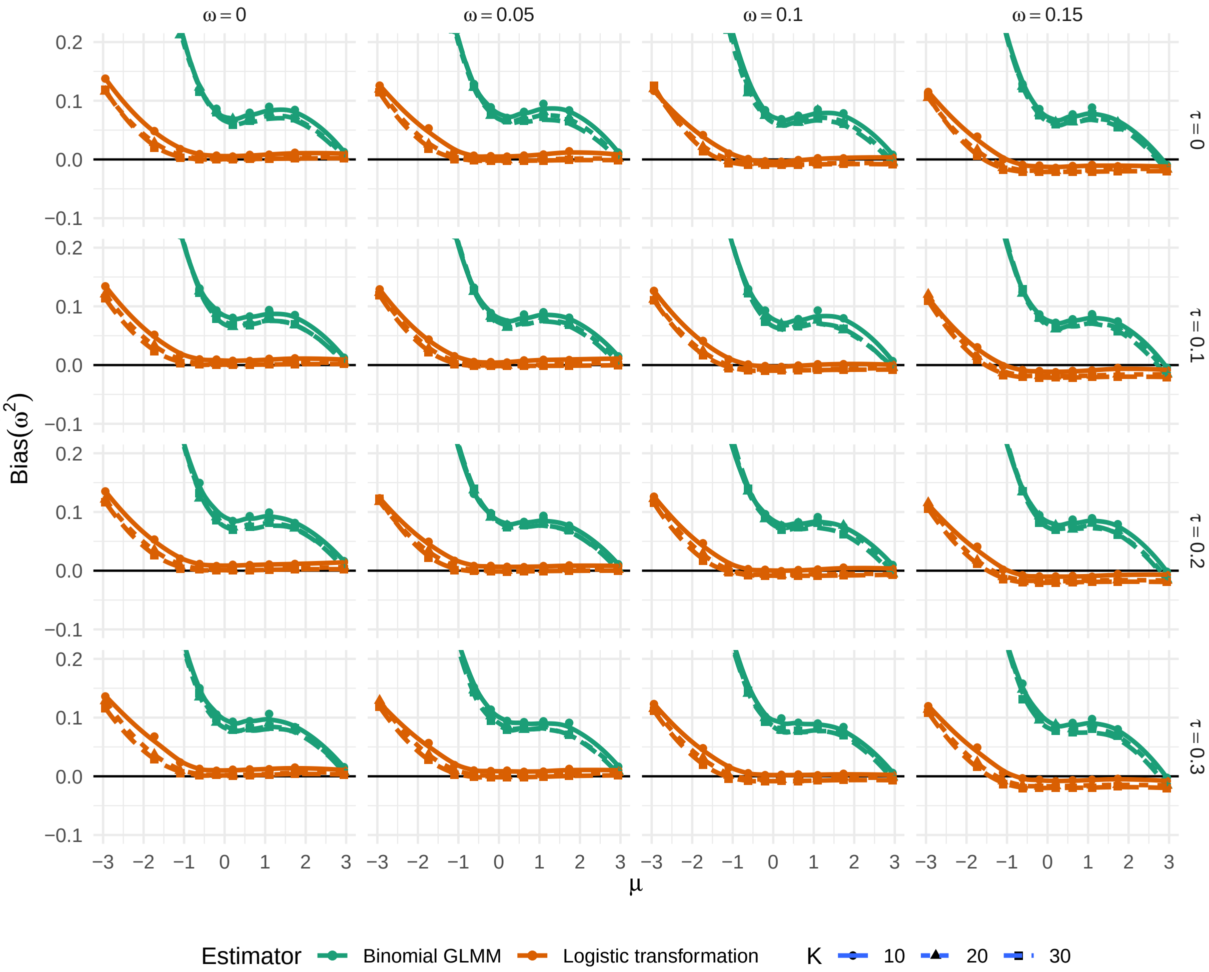

Figure 5

Bias of $\hat{\omega}^2$ using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

the binomial GLMM tends to systematically over-estimate the between-study heterogeneity, with severe bias when the overall average effect size is negative.

Figure 5 depicts the bias of the within-study heterogeneity variance estimator ($\hat{\omega}^2$). The pattern of biases is generally similar to the pattern of for the between-study heterogeneity variance estimator. The estimator based on logistic transformation is close to unbiased except when the overall average effect size is negative. In contrast, the estimator based on the binomial GLMM has a large positive bias except when the overall average effect size is very large and positive. For negative average effect sizes, the binomial GLMM

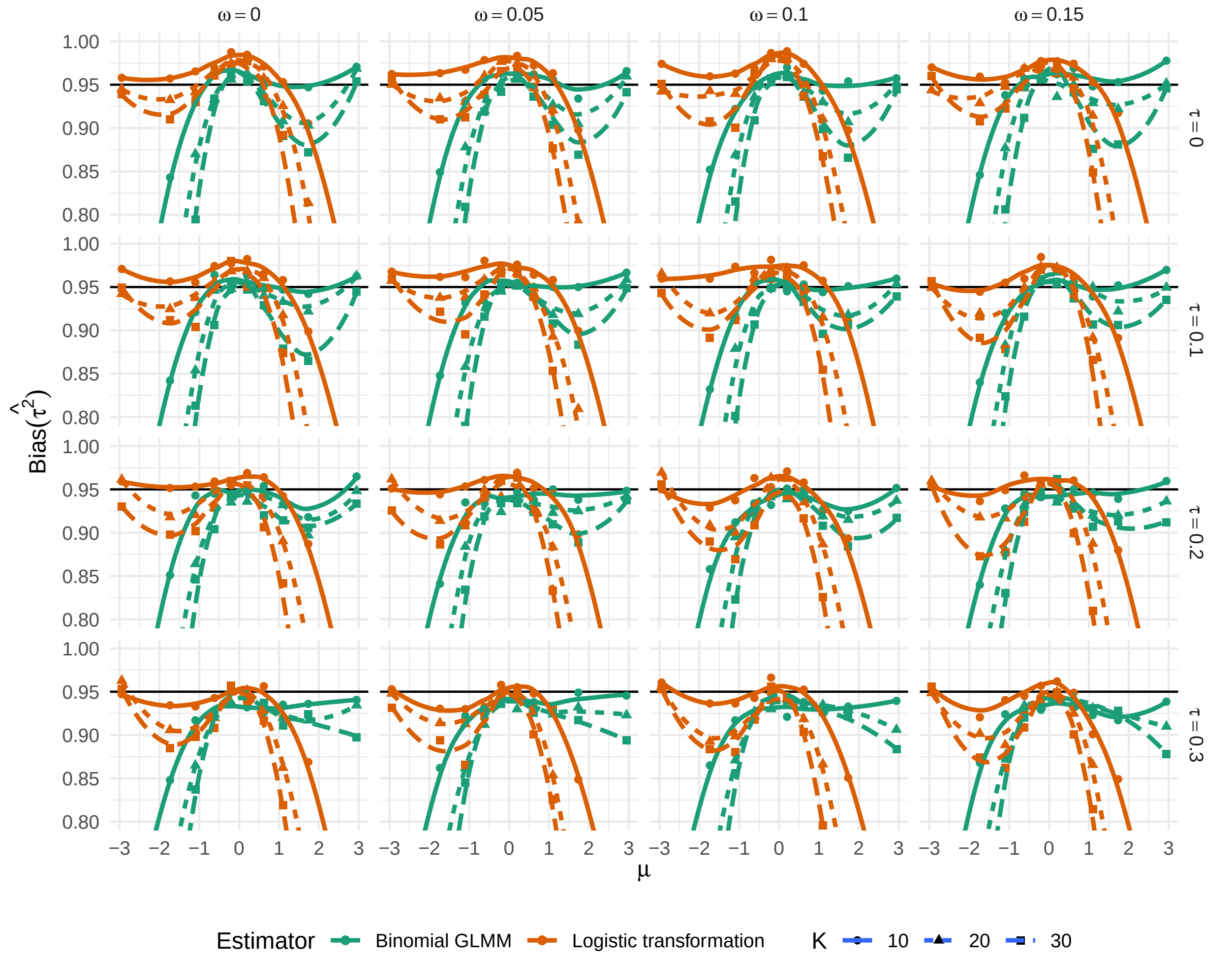

Figure 6

Coverage rates of nominal 95% confidence intervals for μ using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

estimator has a very severe, positive bias.

Figure 6 depicts the empirical coverage rates of 95% confidence intervals for μ based on each of the estimation approaches. Neither approach provides confidence intervals with adequate coverage levels. Because both approaches have systematically biased estimators of μ, confidence intervals are not correctly centered. In regions of the parameter space where the overall average effect size estimator $\hat{\mu}$ is more strongly bias, the corresponding confidence intervals have coverage far below the nominal level. Further, coverage worsens

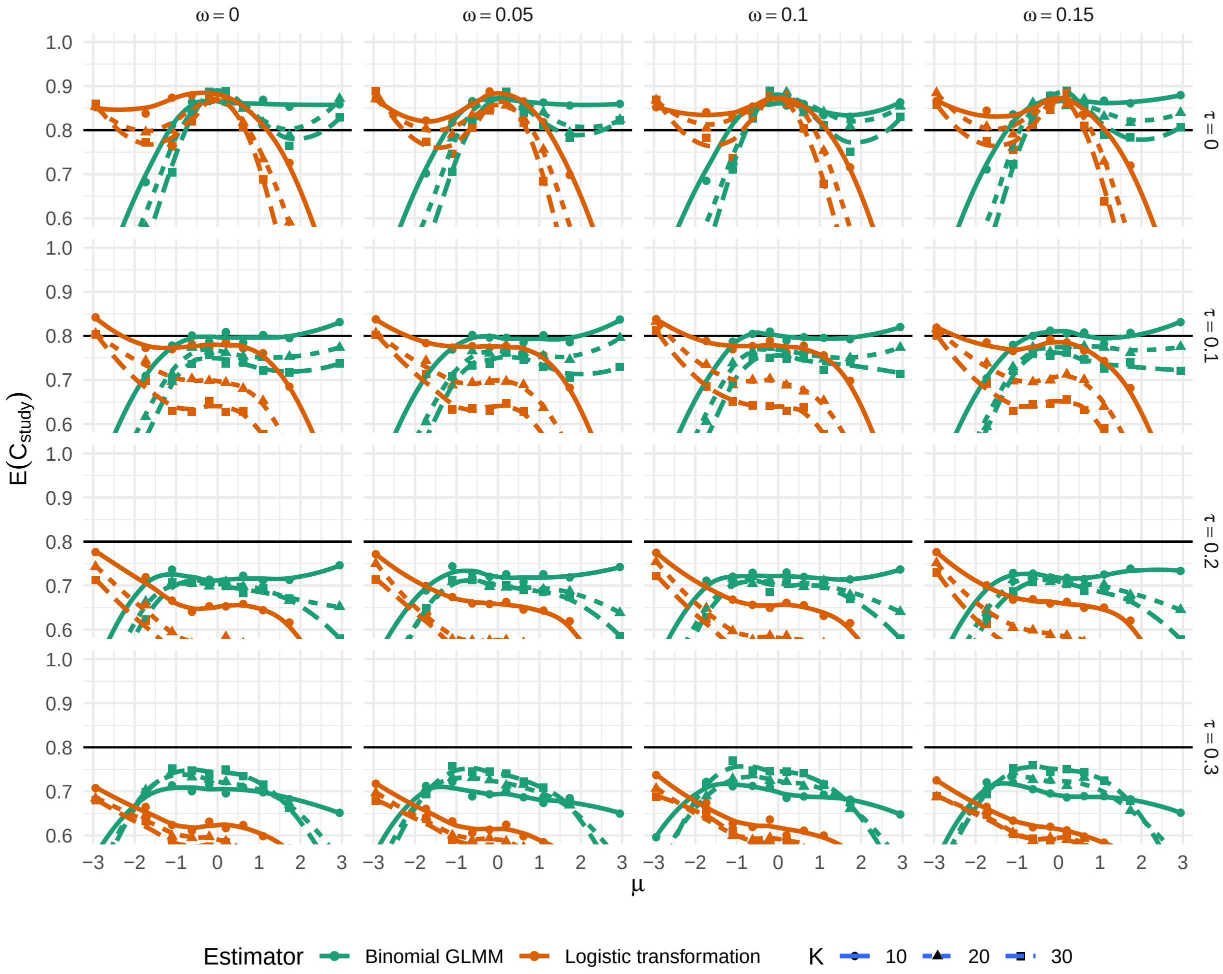

Figure 7

Expected content of nominal 80% study-level prediction intervals using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

as the number of studies increases. This occurs because bias remains stable while sampling variance decreases as K increases.

Figures 7 and 8 depict the expected content of 80% prediction intervals for study-level and case-level effect sizes, respectively. Neither estimation approach yields prediction intervals with expected content near the nominal level. Generally, the pattern of performance is similar to that of the confidence interval coverage levels: with combinations of parameter values where the average effect size estimator $\hat{\mu}$ is systematically biased, the

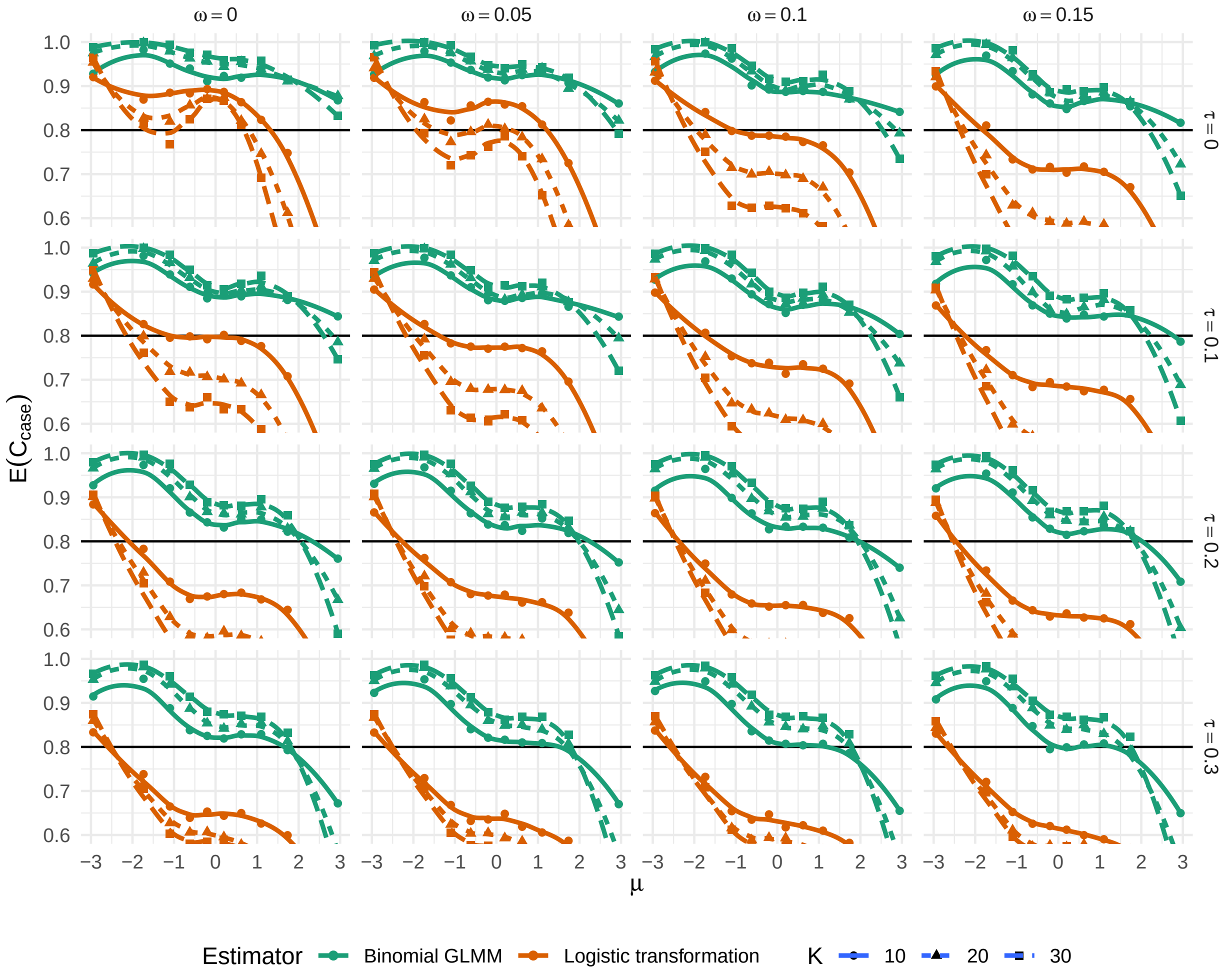

Figure 8

Expected content of nominal 80% case-level prediction intervals using the logistic transformation approach and binomial GLMM approach, for varying values of μ, τ, ω, and K, with Poisson-distributed outcomes.

prediction intervals are not centered correctly and therefore have less-than-adequate expected content. The main exception to the pattern is that the case-level prediction intervals based on the binomial GLMM have above-nominal content over most of the parameter space, which occurs because the binomial GLMM estimator for within-study heterogeneity has such drastic upward bias. Overall, neither method provides prediction intervals that are well calibrated.

Discussion

We have investigated two methods for synthesizing NAP effect sizes, using a meta-analysis model that describes the distribution of effect size parameters on a logistic scale. The approaches involve different distributional approximations. The logistic transformation approach is simple to implement and can be estimated using conventional meta-analysis software. It relies on a delta-method variance approximation, which might not work adequately when the number of observations used to estimate each effect size is small, and it requires truncation of effect size estimates near 0 and 1. The binomial GLMM approach avoids the need for truncation but requires estimating an effective number of trials for the binomial distributional approximation. Again, when only a small number of observations are available to estimate the effect size and its variance, the effective number of trials may not be adequately estimated.

We designed Monte Carlo simulations to emulate realistic conditions for single-case data, including for the number of cases per study and number of observations in baseline and intervention phases. Under these conditions, neither the logistic transformation approach or the binomial GLMM approach work adequately across the full parameter space. The approaches lead to average effect size estimators with different, distinct patterns of bias. This bias, in turn, leads to confidence intervals and prediction intervals with inadequate coverage levels or expected content. Clearly, neither estimation method is ready for use in practice.

Further investigation is needed to understand exactly why the approaches have such severe biases. We suspect that it may be due either to the instability within which the sampling variances are estimated or to correlation between the effect size estimator and its sampling variance. Chen & Pustejovsky (2021) found that a multi-level meta-analysis of the raw NAP effect size estimates performed inadequately due to the strong relationship between the effect size estimator and its sampling variance. Something similar may occur here. In ongoing work, we are exploring how to mitigate these issues by partially pooling the sampling variances or effective number of trials.

References

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, *25*(4), 591–602. https://doi.org/10.1002/sim.2256

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Brannick, M. T., French, K. A., Rothstein, H. R., Kiselica, A. M., & Apostoloski, N. (2021). Capturing the underlying distribution in meta-analysis: Credibility and tolerance intervals. *Research Synthesis Methods*, *12*(3), 264–290. https://doi.org/10.1002/jrsm.1479

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://journal.r-project.org/archive/2017/RJ-2017-066/index.html

Chen, M., & Pustejovsky, J. E. (2021). *Multi-level meta-analysis of single-case experimental designs using robust variance estimation* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/59h32

Ganz, J., Pustejovsky, J. E., Reichle, J., Vannest, K., Foster, M., Pierson, L. M., Wattanawongwan, S., Bernal, A., Chen, M., Haas, A. N., Sallese, M. R., Skov, R., & Smith, S. D. (2021). *Participant characteristics predicting communication outcomes in AAC implementation for individuals with ASD and IDD: A systematic review and meta-analysis* [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/6sgba

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

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association*, *85*(411), 793. https://doi.org/10.2307/2290017

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. https://doi.org/10.1016/j.jsp.2013.11.003

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). *Non-overlap analysis for single-case research* (T. R. Kratochwill & J. R. Levin, Eds.; pp. 127–151). American Psychological Association. https://doi.org/10.1037/14376-005

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. https://doi.org/10.1016/j.jsp.2018.02.003

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

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. https://doi.org/10.1037/met0000179

Pustejovsky, J. E., & Ferron, J. (2017). Research synthesis and meta-analysis of single-case designs. In *Handbook of Special Education* (2nd Edition, p. 63). Routledge.

Ryu, E., & Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, *27*(10), 1703–1717. https://doi.org/10.1002/sim.3079

Sen, P. K. (1967). A note on asymptotically distribution-free confidence bounds for P{X < Y}, based on two independent samples. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *29*(1), 95–102.

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

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.2307/1165329

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03