eAppendix: Robust Metrics and Sensitivity Analyses for Meta-Analyses of Heterogeneous Effects

Contents

1 Theory **3**

1.1 Sign test method 3

1.2 Sensitivity analysis for unmeasured confounding 4

1.2.1 Setting and notation 4

1.2.2 Proportion of studies with scientifically meaningful effect sizes as a function of the bias factor 5

1.2.3 Bias factor or confounding strength required to reduce proportion of scientifically meaningful effect sizes to below a threshold 6

1.2.4 Meta-analyses including both randomized and observational studies . 7

2 Simulation study **7**

2.1 Methods 7

2.2 Summary of results 9

2.3 All coverage results by distribution 12

2.4 All confidence interval width results by distribution 16

2.5 All RMSE results by distribution 20

2.6 All bias results by distribution 24

3 Applied example **28**

4 Software **31**

1. Theory

1.1. Sign test method

In addition to the calibrated estimation method described in the main text, we considered an additional robust method that follows straightforwardly from repurposing existing methods for conducting inference on the percentiles of a heterogeneous effect distribution[1]. Let k denote the number of studies in the meta-analysis and let $\widehat{B}_i = 1\{\widehat{\theta}_i < q\} - 1\{\widehat{\theta}_i > q\}$. Wang et al. (2010) proposed the test statistic[1]:

$$\widehat{T}(q) = \sum_{i=1}^{k} \left| \Phi\left((q - \widehat{\theta}_i)/\widehat{\sigma}_i\right) - 1/2 \right| \widehat{B}_i = \sum_{i=1}^{k} \left\{ \Phi\left((q - \widehat{\theta}_i)/\widehat{\sigma}_i\right) - 1/2 \right\}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. To provide intuition, the term $\Phi\left((q - \widehat{\theta}_i)/\widehat{\sigma}_i\right)$ represents the asymptotic coverage level of the confidence interval $[-\infty, q]$ for θ_i, that is, $P(q > \theta_i)$.[a] Thus, the term in the absolute value serves as a precision weight for the contribution $\widehat{B}_i$ in that it compares the precision of study i (specifically with respect to the threshold q) to that of a study that is maximally uninformative in the sense that $P(q > \theta_i) = 1/2$. To test a null hypothesis equivalent to $H_0 : 1 - P_{>q} = p^*$ for a fixed percentile p^*, Wang et al. (2010)[1] simulated a reference distribution for $\widehat{T}(q)$ under H_0, calling the reference test statistic $\widehat{T}^*(q)$:

$$\widehat{T}^*(q) = \sum_{i=1}^{k} \left| \Phi\left((q - \widehat{\theta}_i)/\widehat{\sigma}_i\right) - 1/2 \right| \Delta_i$$

where Δ_i is a null counterpart to $\widehat{B}_i$ that is simulated to equal 1 with probability p^* and to equal -1 with probability $1 - p^*$. Wang et al. (2010)[2] showed that this simulated distribution approximates the true distribution under H_0 when the sample size in the meta-analyzed studies is large, though without requiring asymptotics on the number of meta-analyzed studies (k). These results designed for conducting inference on a fixed percentile of interest also allow straightforward inference and point estimation for the proportion of effects above a threshold, $P_{>q}$. To do so, one can specify a grid of M values $(p_1^*, \cdots, p_M^*)$ ranging from 0

[a] Specifically, let n_i be the sample size in the i^{th} study. Then $\widehat{\theta}_i \xrightarrow[n_i \to \infty]{D} N\left(\theta_i, \sigma_i^2\right)$. Consider the coverage of a confidence interval for $\widehat{\theta}_i$ with lower bound $\widehat{\theta}_i - c\widehat{\sigma}_i$ for an arbitrary constant $c > 0$. Asymptotically, the probability that the lower bound is too high to cover θ_i is $P\left(\widehat{\theta}_i - c\widehat{\sigma}_i > \theta_i\right) \xrightarrow[n_i \to \infty]{p} \Phi(-c)$. Setting $q = \widehat{\theta}_i - c\widehat{\sigma}_i$ yields $P(q > \theta_i) = \Phi\left((q - \widehat{\theta}_i)/\widehat{\sigma}_i\right)$, as desired.

to 1 and conduct a level-α test of each hypothesis $H_{0,m} : 1 - P_{>q} = p^*_m$ by simulating many iterates of the reference statistics $\widehat{T}^*(q)$, whose distribution depends on p^* via the random binary variable Δ_i. The set of $1 - p^*_m$ that are not rejected at level α form the $100 \times (1 - \alpha)\%$ confidence interval for $\widehat{P}_{>q}$, and a point estimate $\widehat{P}_{>q}$ can be defined as the value $1 - p^*_m$ with the largest p-value (which we term the "sign test max" in the simulation study).

1.2. Sensitivity analysis for unmeasured confounding

We previously developed sensitivity analyses for unmeasured confounding in meta-analyses; these methods quantified the proportion of effects of scientifically meaningful magnitude, $\widehat{P}_{>q}$, under a specified amount of unmeasured confounding[3]. We also developed converse methods to estimate the strength of confounding capable of reducing $\widehat{P}_{>q}$ itself to below a chosen threshold[3]. These methods used parametric point estimation and inference that generalized the parametric methods described in the main text here[4]. These sensitivity analysis methods can be conducted robustly using the present calibration-based methods as follows.

1.2.1 Setting and notation

This background material is partly reproduced from our previous work regarding sensitivity analysis using parametric methods, where we provide more detail, intuition, and guidance on practical interpretation[5,3]. Let X denote a binary exposure, Y a binary outcome, Z a vector of measured confounders, and U one or more unmeasured confounders[5]. Consider the point estimate for a single meta-analyzed study on the relative risk scale; other effect size measures, such as standardized mean differences and odds ratios, can be approximately converted to relative risks to allow application of these methods, as when conducting the sensitivity analyses parametrically[3]. Let:

$$RR_{XY|z} = \frac{P\left(Y = 1 \mid X = 1, Z = z\right)}{P\left(Y = 1 \mid X = 0, Z = z\right)}$$

be the confounded relative risk (RR) of Y for $X = 1$ versus $X = 0$ conditional or stratified on the measured confounders $Z = z$. Let its unconfounded counterpart standardized to the population be:

$$RR'_{XY|z} = \frac{\sum_u P\left(Y = 1 \mid X = 1, Z = z, U = u\right) P\left(U = u \mid Z = z\right)}{\sum_u P\left(Y = 1 \mid X = 0, Z = z, U = u\right) P\left(U = u \mid Z = z\right)}$$

Define the ratio of the confounded to the unconfounded relative risks as $B = RR_{XY|z}/RR'_{XY|z}$; this "bias factor" can be sharply bounded as follows[5]. Let:

$$RR_{Xu} = P\left(U = u \mid X = 1\right)/P\left(U = u \mid X = 0\right)$$

Define the first sensitivity parameter as $RR_{XU} = \max_u\left(RR_{Xu}\right)$; that is, the maximal relative risk of $U = u$ for $X = 1$ versus $X = 0$ across strata of U. (If U is binary, this is just the relative risk relating X and U.) Next, for each stratum x of X, define a relative risk of U on Y, maximized across all possible contrasts of U:

$$RR_{UY|X=x} = \frac{\max_u P\left(Y = 1|X = x, U = u\right)}{\min_u P\left(Y = 1|X = x, U = u\right)}, x \in \{0, 1\}$$

Define the second sensitivity parameter as $RR_{UY} = \max\left(RR_{UY|X=0}, RR_{UY|X=1}\right)$. That is, considering both strata of X, it is the largest of the maximal relative risks of U on Y conditional on X. Others[5] showed that when $B \geq 1$, then B itself is bounded above by:

$$B \leq \frac{RR_{XU} \cdot RR_{UY}}{RR_{XU} + RR_{UY} - 1}$$

and that when $B \leq 1$, the same bound holds for $1/B$. Thus, defining the "worst-case" bias factor as $B^+ = \frac{RR_{XU} \cdot RR_{UY}}{RR_{XU} + RR_{UY} - 1}$, a sharp bound for the unconfounded effect when the observed $RR_{XY|z} \geq 1$ is $RR'_{XY|z} \geq RR_{XY|z}/B^+$, and a sharp bound when $RR_{XY|z} \leq 1$ is $RR'_{XY|z} \leq RR_{XY|z} \times B^+$.

1.2.2 Proportion of studies with scientifically meaningful effect sizes as a function of the bias factor

Here, we consider the case in which B is assumed to be homogeneous across studies; that is, all studies are subject to the same degree of unmeasured confounding, albeit possibly due to different unmeasured confounders. Assuming homogeneous, rather than heterogeneous, bias across studies yields conservative sensitivity analyses in some settings; see Mathur & VanderWeele's (2019)[3] Table 1 for details.

As one metric of sensitivity to unmeasured confounding, we can estimate the proportion of unconfounded effects stronger than q when all studies have bias factor B (i.e., all studies' relative risk estimates are shifted away from the null by a factor of B due to unmeasured confounding). This quantity, here denoted $\widehat{P}_{>q}\left(B\right)$, can be robustly estimated using the

calibrated estimates as follows. First, for a chosen bias factor B, define a bias-corrected point estimate for each study on the log relative risk scale as $\widehat{\theta}_i' = \log\left(RR_{XY|z}/B\right) = \widehat{\theta}_i - \log B$ if $\widehat{\mu} > 0$ (i.e., the confounded pooled point estimate on the log relative risk scale is apparently causative) or as $\widehat{\theta}_i' = \log\left(RR_{XY|z} \times B\right) = \widehat{\theta}_i + \log B$ if $\widehat{\mu} < 0$ (i.e., the confounded pooled point estimate is apparently preventive). Similarly, define the bias-corrected pooled point estimate $\widehat{\mu}' = \widehat{\mu} - \log B$ if $\widehat{\mu} > 0$ and as $\widehat{\mu}' = \widehat{\mu} + \log B$ if $\widehat{\mu} < 0$. Next, use the bias-corrected point estimates $\widehat{\theta}_i'$ and pooled point estimate $\widehat{\mu}'$ to calculate bias-corrected calibrated estimates on the log relative risk scale as $\widetilde{\theta}_i' = \widehat{\mu}' + \sqrt{\widehat{\tau}^2/(\widehat{\tau}^2 + \widehat{\sigma}_i^2)}\left(\widehat{\theta}_i' - \widehat{\mu}'\right)$. (Note that because B is constant across studies, no bias correction is needed for $\widehat{\tau}^2$.) Then $\widehat{P}_{>q}(B)$ can be straightforwardly estimated using the sample proportion of bias-corrected point estimates that are stronger than q, i.e., $\widehat{P}_{>q}(B) = \widehat{P}\left(\widetilde{\theta}_i' > q\right)$. A confidence interval can be obtained via bias-corrected and accelerated (BCa) bootstrapping as described in the main text, resampling the pairs $(\widehat{\theta}_i', \widehat{\sigma}_i)$.

1.2.3 Bias factor or confounding strength required to reduce proportion of scientifically meaningful effect sizes to below a threshold

As a second sensitivity analysis metric, we previously proposed reporting the minimum bias factor in each study that would be required to reduce to less than r (e.g., 0.10) the proportion of effects of scientifically meaningful magnitude[3]. This metric, denoted $\widehat{T}(r, q)$, can be robustly estimated using a simple grid search across values of B. That is, for each of a grid of candidate values for the bias factor B, such as $(1, 1.01, 1.02, ...)$, apply the methods described in Section 1.2.2 to estimate $\widehat{P}_{>q}(B)$. Then, $\widehat{T}(r, q)$ is simply the bias factor such that $\widehat{P}_{>q}(B)$ is exactly equal to the chosen proportion r; that is, $\widehat{T}(r, q) = B \ : \ \widehat{P}_{>q}(B) = r$.

Recall that B^+, the upper bound on B, is a function of two sensitivity parameters that characterize the strengths of association between the unmeasured confounder(s) and the exposure (RR_{XU}) and between the unmeasured confounder(s) and the outcome (RR_{UY}). If these two sensitivity parameters are assumed to be equal to one another[6,3], the sensitivity analysis metric $\widehat{T}(r, q)$ can alternatively be parameterized on the more intuitive confounding strength scale (i.e., the values of both RR_{XU} and RR_{UY}). Consider the minimum confounding strength required to reduce to less than r the proportion of studies with scientifically meaningful effect sizes. This quantity, denoted $\widehat{G}(r, q)$, can be obtained as a simple transformation of $\widehat{T}(r, q)$ as follows: $\widehat{G}(r, q) = \widehat{T}(r, q) + \sqrt{\left(\widehat{T}(r, q)\right)^2 - \widehat{T}(r, q)}$. This metric is closely analogous to the "E-value" for an individual study[6,5].

Confidence intervals for $\widehat{T}(r,q)$ or $\widehat{G}(r,q)$ can be constructed via bootstrapping by resampling the bias-corrected point estimates, $\widehat{\theta}_i'$, estimating the desired quantity ($\widehat{T}(r,q)$ or $\widehat{G}(r,q)$) for each resample, and constructing confidence intervals using the BCa method. In practice, the same set of resamples could be used for $\widehat{P}_{>q}(B)$, $\widehat{T}(r,q)$, and $\widehat{G}(r,q)$.

1.2.4 Meta-analyses including both randomized and observational studies

The present robust methods also allow straightforward extension to meta-analyses in which some studies (e.g., randomized studies) are assumed to have no unmeasured confounding, while other studies are assumed to be subject to unmeasured confounding of strength B. For this setting, set the bias-corrected point estimates $\widehat{\theta}_i'$ equal to their observed values ($\widehat{\theta}_i$) for the randomized studies and to $\widehat{\theta}_i - \log B$ or $\widehat{\theta}_i + \log B$, as described above, for the observational studies. Then, meta-analyze these bias-corrected point estimates to arrive at a bias-corrected pooled point estimate and heterogeneity estimate, $\widehat{\mu}'$ and $\widehat{\tau}^{2\prime}$, and use these to compute the calibrated estimates. (Note that because B is no longer constant across all studies, it is now necessary to estimate $\widehat{\tau}^{2\prime}$ using the calibrated, bias-corrected estimates.) Estimation and inference for $\widehat{P}_{>q}(B)$, $\widehat{T}(r,q)$, and $\widehat{G}(r,q)$ would then proceed as above.

2. Simulation study

2.1. Methods

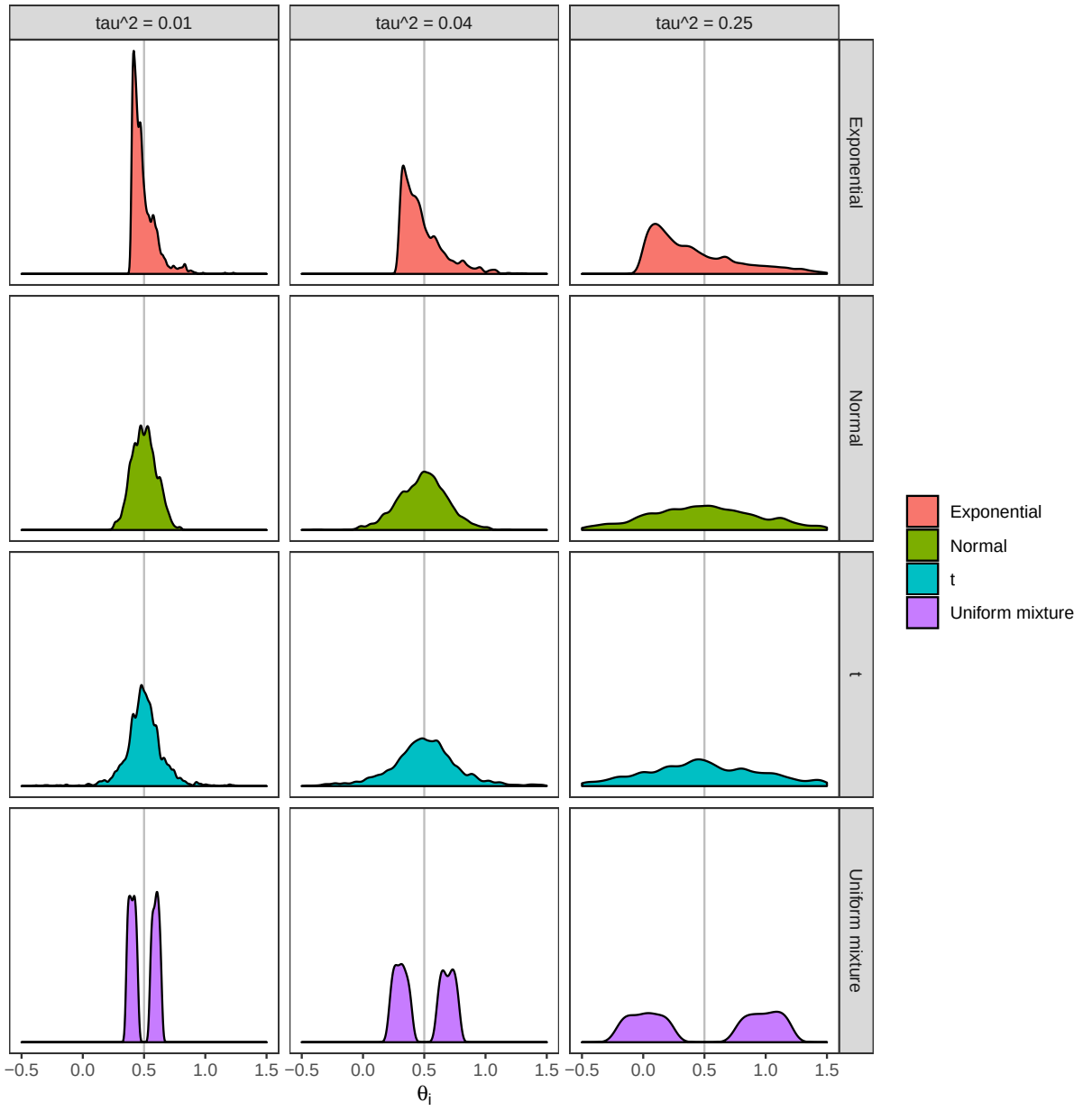
The simulation study assessed the performance of point estimation and inference methods for $\widehat{P}_{>q}$ without confounding. We fixed the mean of the true population effects to $\mu = 0.50$ on the mean difference scale while varying the number of studies (k) between 5 and 50, the heterogeneity (the variance of true population effects) $\tau^2 \in \{0.01, 0.04, 0.25\}$, and the distribution of total sample sizes within each study (either $N \sim \text{Unif}(100, 200)$ or $N \sim \text{Unif}(800, 900)$). For each of k meta-analyzed studies, we generated a true effect, θ_i, on the raw mean difference scale from a normal distribution, a scaled and shifted t-distribution with 3 degrees of freedom, a bimodal uniform mixture distribution, or a shifted exponential distribution. For all distributions, we chose the parameters to provide the desired mean of $\mu = 0.50$ and heterogeneity τ^2. Figure 1 shows example data depicting true population effects simulated from each of the four distributions for each value of τ^2.

We then simulated subject-level data for a control group with mean 0 and for a treatment with mean θ_i; each group was of size $N/2$ with a standard deviation of 1. Thus, the within-study standard error of the estimated mean difference, $\widehat{\theta}_i$, was approximately $\widehat{\sigma}_i = \sqrt{4/N}$. For the meta-analysis, the proportion of the total variance attributable to effect heterogeneity[7,8] (termed I^2) was approximately $\tau^2/\left(\tau^2 + 4/E[N]\right)$. We chose values of q to result in true proportions $P_{>q}$ of 0.05, 0.10, 0.20, and 0.50.

We ran scenarios representing all 480 possible combinations of the varying parameters, using 5,000 iterates to estimate both types of bootstrap confidence interval and 2,000 iterates to estimate the reference distribution for the sign test method[1]. We ran at least 500 simulation iterates per scenario. For inference, we assessed the coverage and width of 95% confidence intervals constructed by computing $\widehat{P}_{>q}$ from calibrated estimates in bootstrapped datasets as described above ("BCa-calibrated"), constructed using the delta method ("Parametric")[4], and constructed by estimating $\widehat{P}_{>q}$ parametrically in bootstrapped datasets ("BCa-parametric") as we previously described[4]. For point estimation, we assessed the root mean squared error (RMSE) and absolute bias of three methods: the parametric method ("Parametric"), the sample proportion based on the calibrated estimates ("Calibrated"), and the value maximizing the p-value of the sign test as described above ("Sign test max.").

	$\tau^2 = 0.01$	$\tau^2 = 0.04$	$\tau^2 = 0.25$
$E[N] = 150$	0.27	0.60	0.90
$E[N] = 850$	0.68	0.89	0.98

eTable 1: *Approximate values of relative heterogeneity (I^2) for each combination of simulation parameters regarding the mean within-study sample size ($E[N]$) and heterogeneity (τ^2).*



eFigure 1: *Distributions of true population effects (θ_i) used in simulation study for varying choices of heterogeneity*

2.2. Summary of results

Tables 2 and 3 summarize the performance of 95% confidence intervals and of point estimates respectively. For clarity given the very large number of simulation scenarios, the tables summarize results according to a single varied simulation parameter that produced particularly interesting variation in performance across methods (i.e., τ^2 for inference and the true effect

distribution for point estimation). Comprehensive simulation results for all 480 scenarios are also presented in the figures of Sections 2.3–2.6; these results are also publicly available as a dataset (`https://osf.io/6nyg8/`).

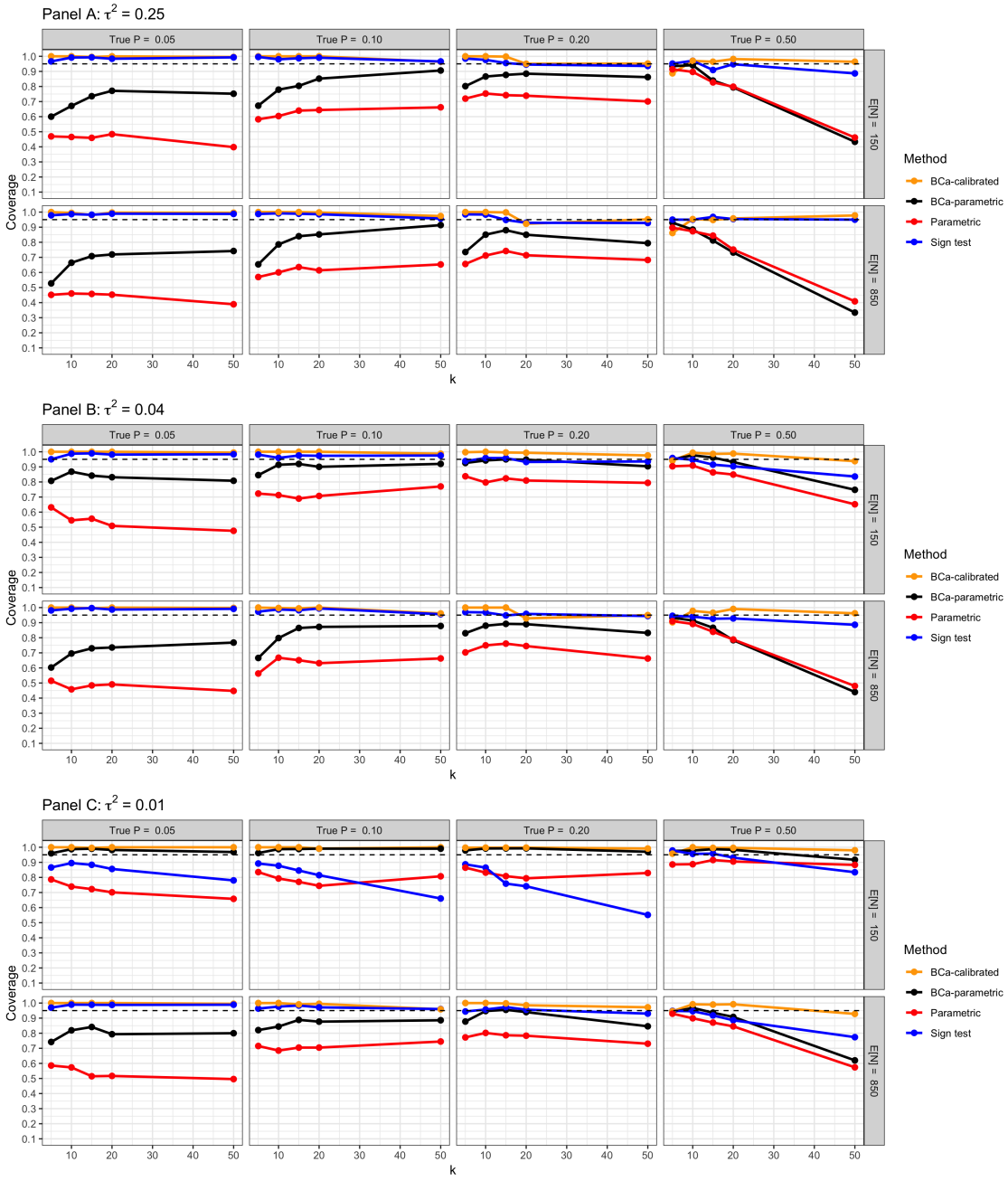
τ^2	**Method**	**Coverage**	**Minimum coverage**	**Width**
0.01	BCa-calibrated	0.99	0.90	0.52
	BCa-parametric	0.94	0.62	0.50
	Parametric	0.84	0.50	0.36
	Sign test	0.90	0.16	0.37
0.04	BCa-calibrated	0.98	0.91	0.38
	BCa-parametric	0.89	0.44	0.35
	Parametric	0.82	0.45	0.28
	Sign test	0.96	0.72	0.33
0.25	BCa-calibrated	0.98	0.92	0.32
	BCa-parametric	0.86	0.33	0.28
	Parametric	0.82	0.39	0.25
	Sign test	0.97	0.84	0.31

eTable 2: *Performance of 95% confidence intervals for scenarios with $k \geq 10$, showing mean coverage, minimum coverage, and mean width of 95% confidence intervals aggregating all scenarios with a given amount of heterogeneity (τ^2).*

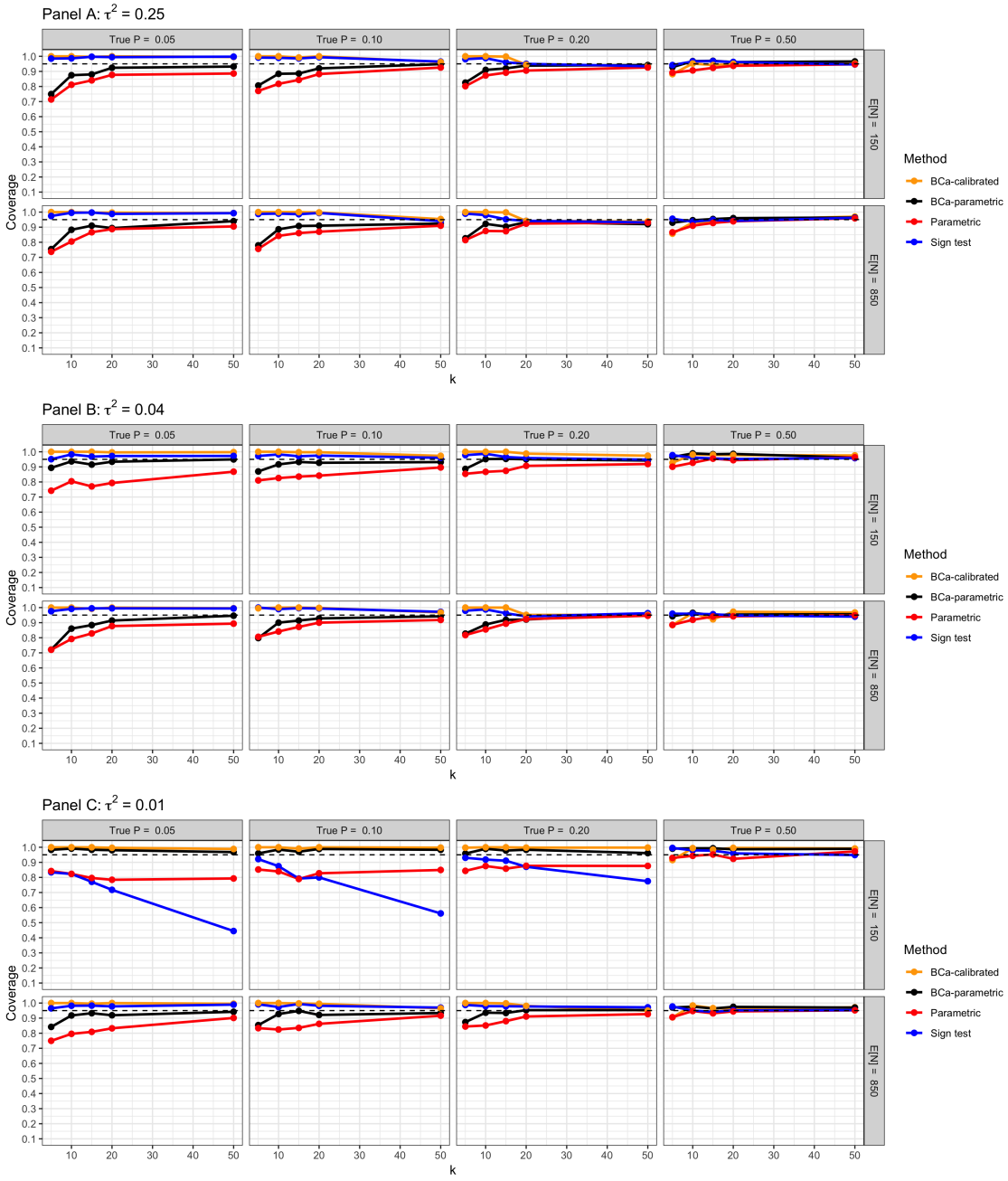
Distribution	Method	RMSE	Bias	Absolute bias
Exponential	Calibrated	0.162	0.022	0.097
	Parametric	0.164	0.042	0.105
	Sign test max.	0.141	0.036	0.089
Normal	Calibrated	0.159	0.005	0.094
	Parametric	0.135	0.006	0.080
	Sign test max.	0.135	0.014	0.084
t	Calibrated	0.151	0.007	0.089
	Parametric	0.150	0.027	0.088
	Sign test max.	0.135	0.010	0.082
Uniform mixture	Calibrated	0.165	0.010	0.098
	Parametric	0.135	0.007	0.083
	Sign test max.	0.137	0.015	0.087

eTable 3: *Performance of methods for point estimation, showing means across all scenarios for each distribution of root mean squared error (RMSE), bias, and absolute bias.*

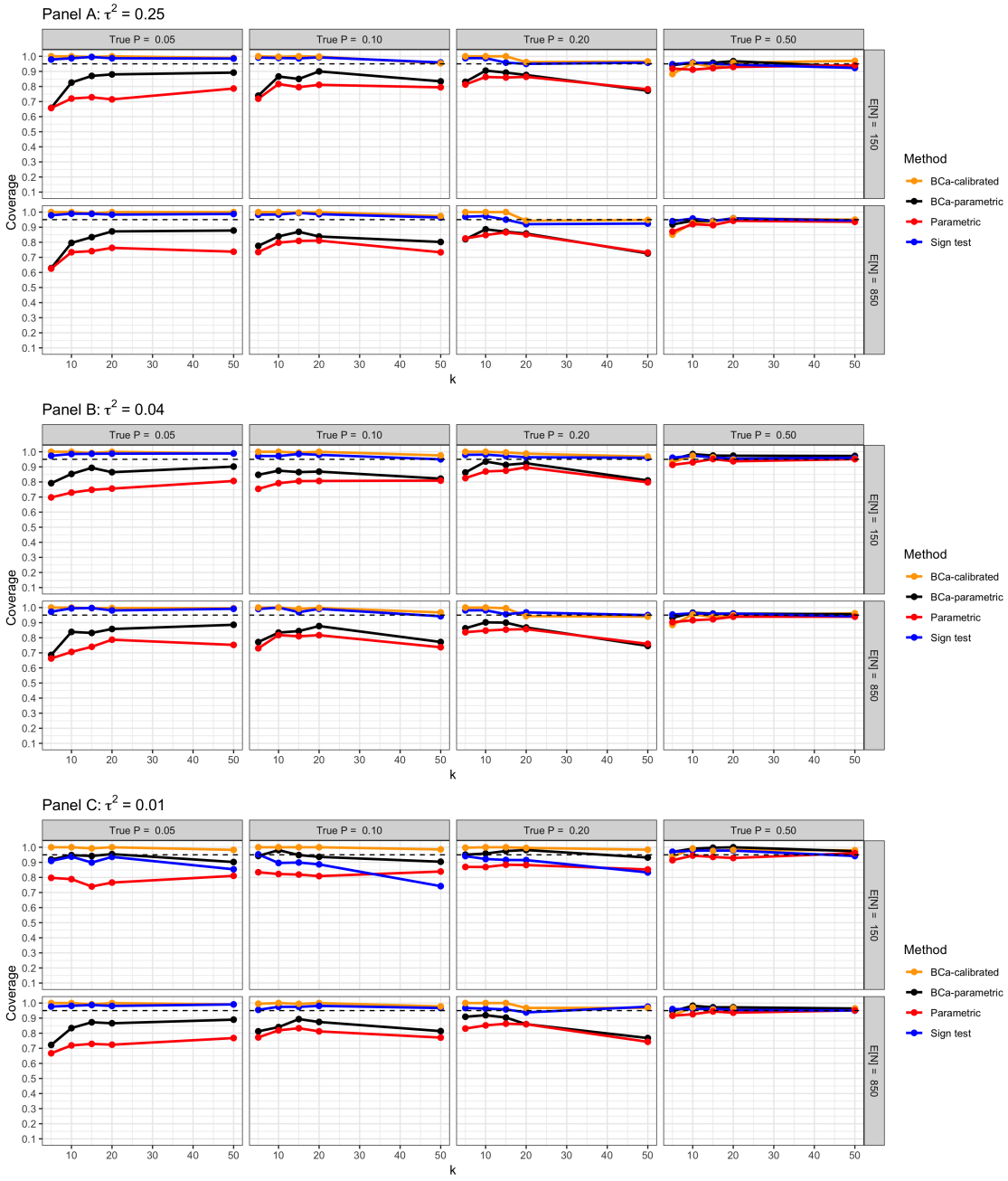
2.3. All coverage results by distribution



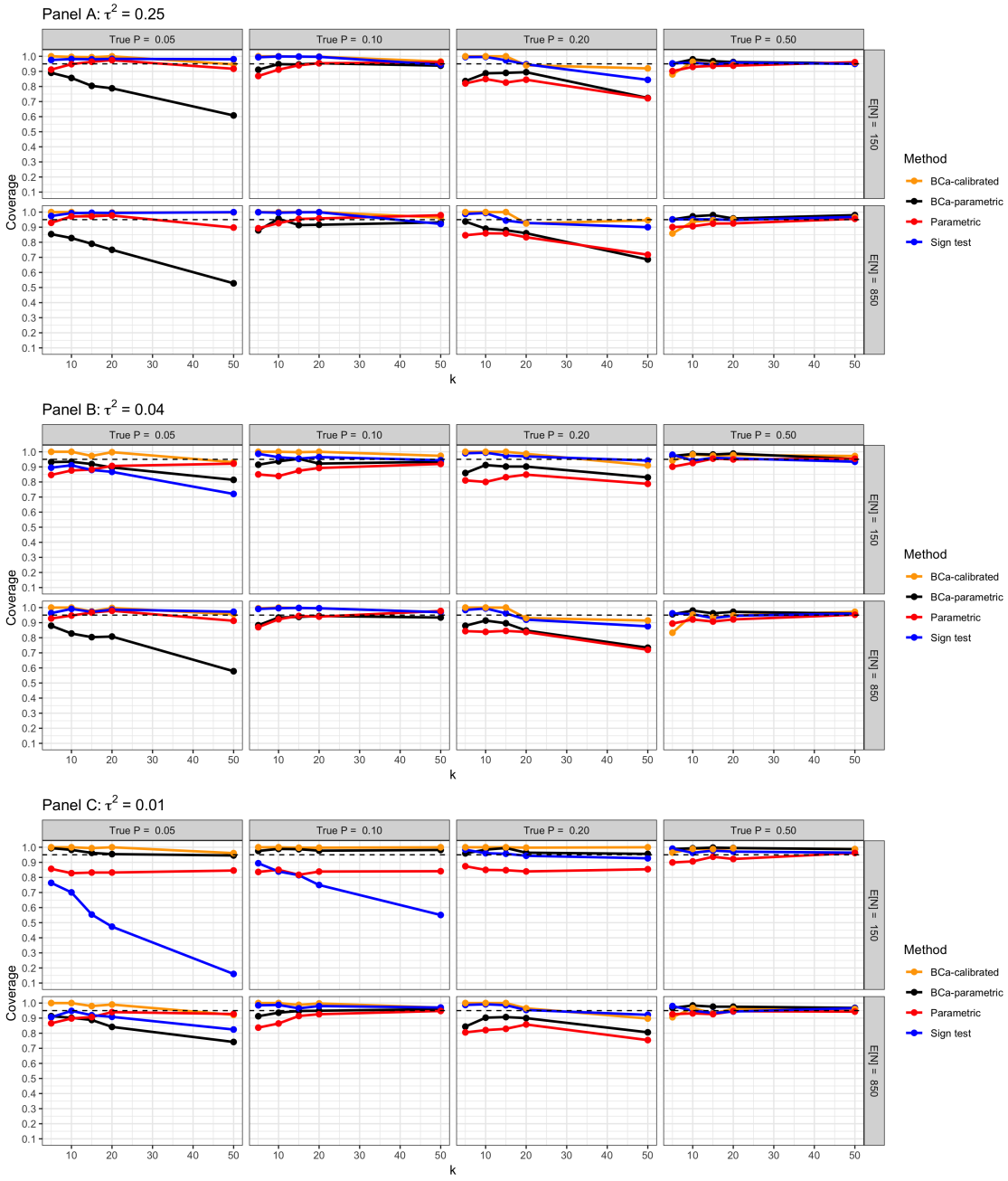
eFigure 2: *Coverage of 95% confidence intervals for exponential distribution*



eFigure 3: *Coverage of 95% confidence intervals for normal distribution*

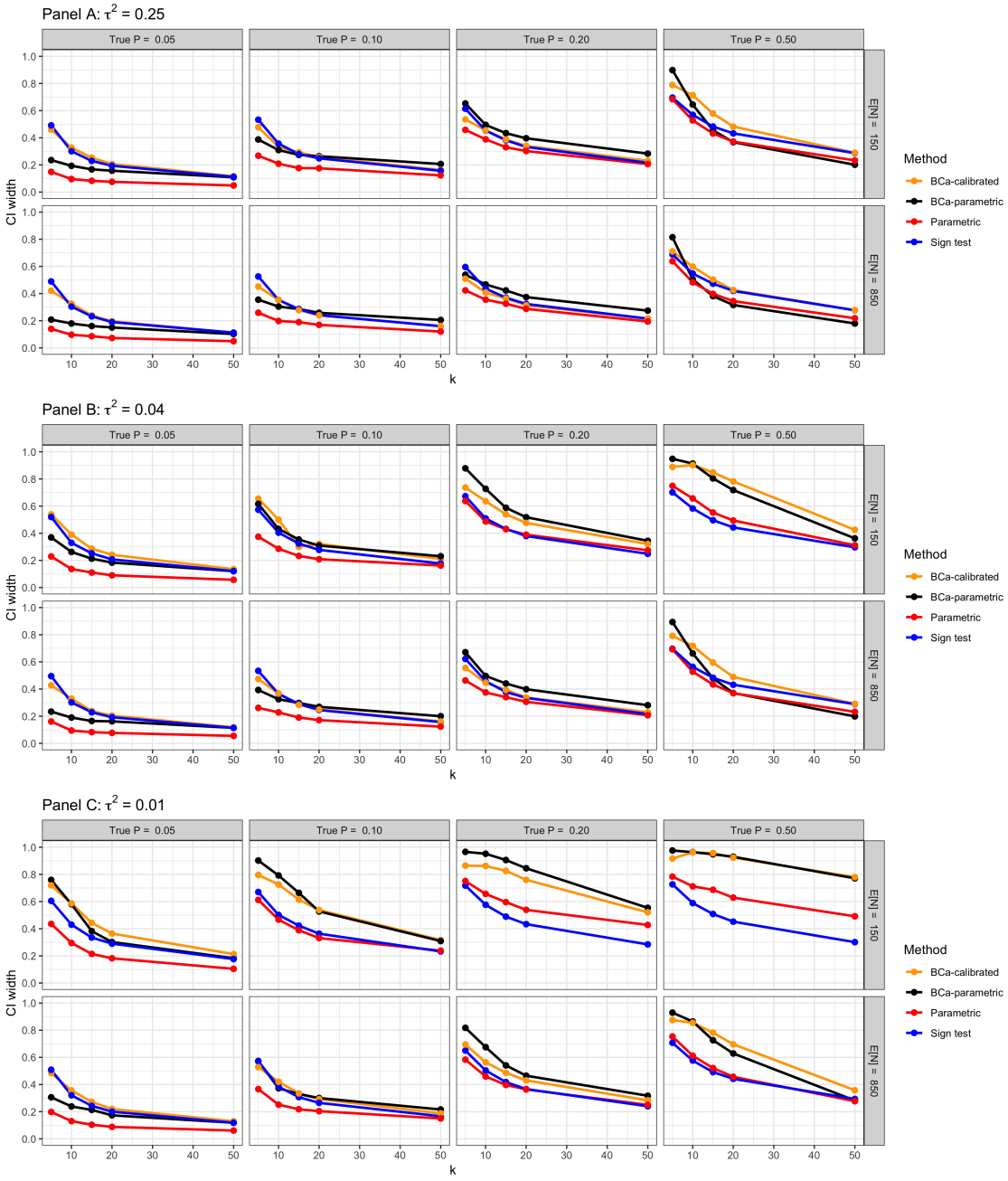


eFigure 4: *Coverage of 95% confidence intervals for scaled t distribution*

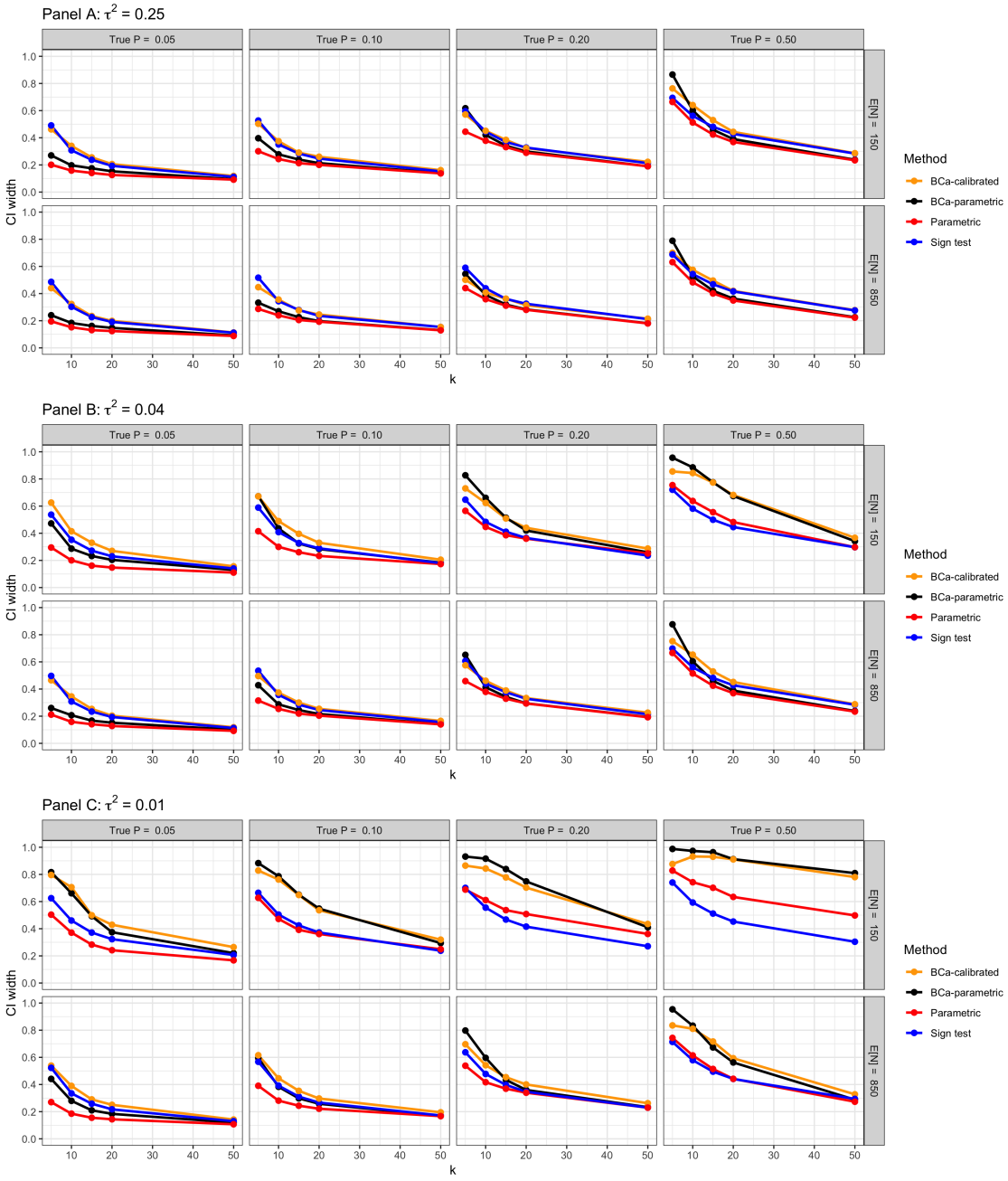


eFigure 5: *Coverage of 95% confidence intervals for uniform mixture distribution*

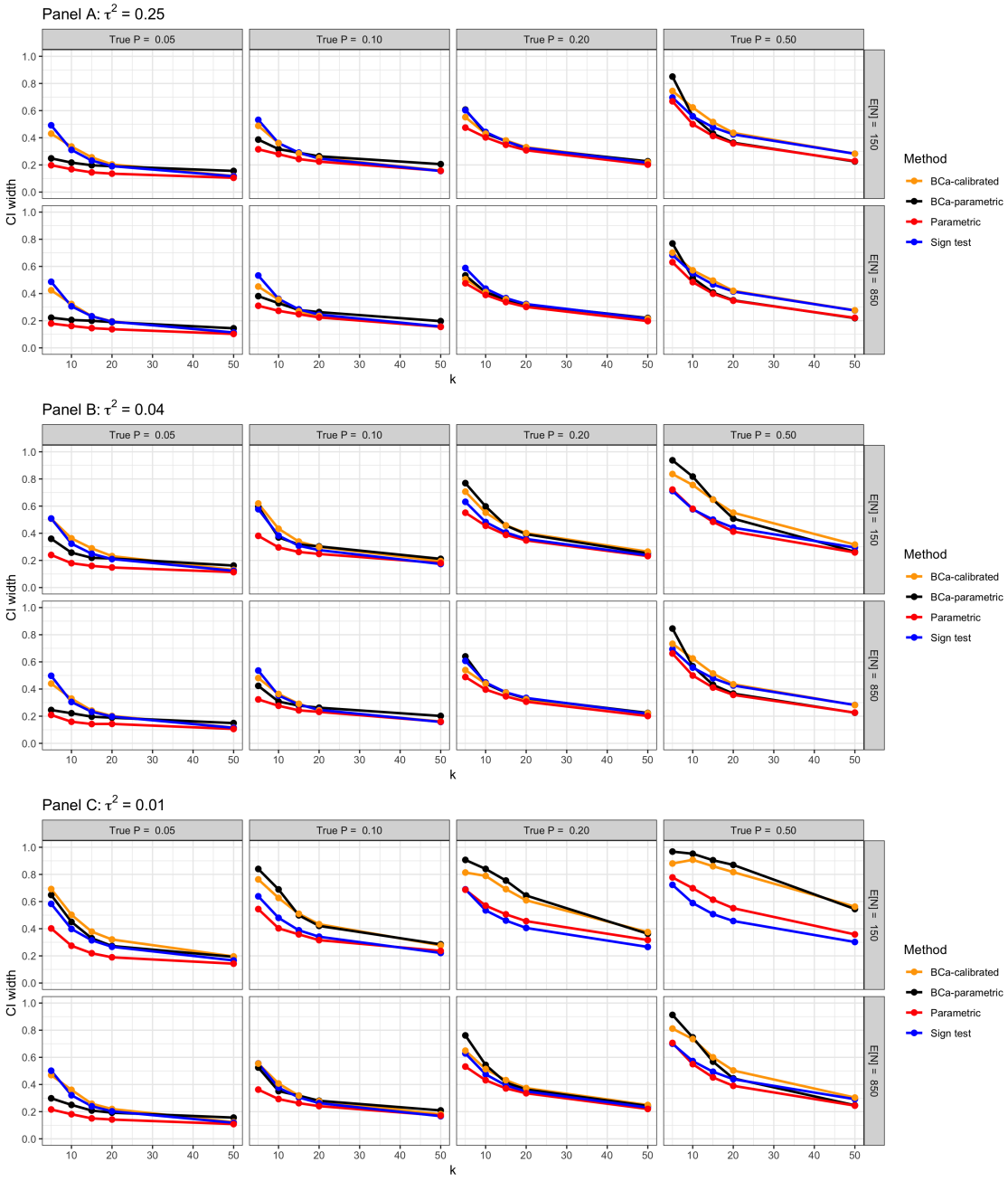
2.4. All confidence interval width results by distribution



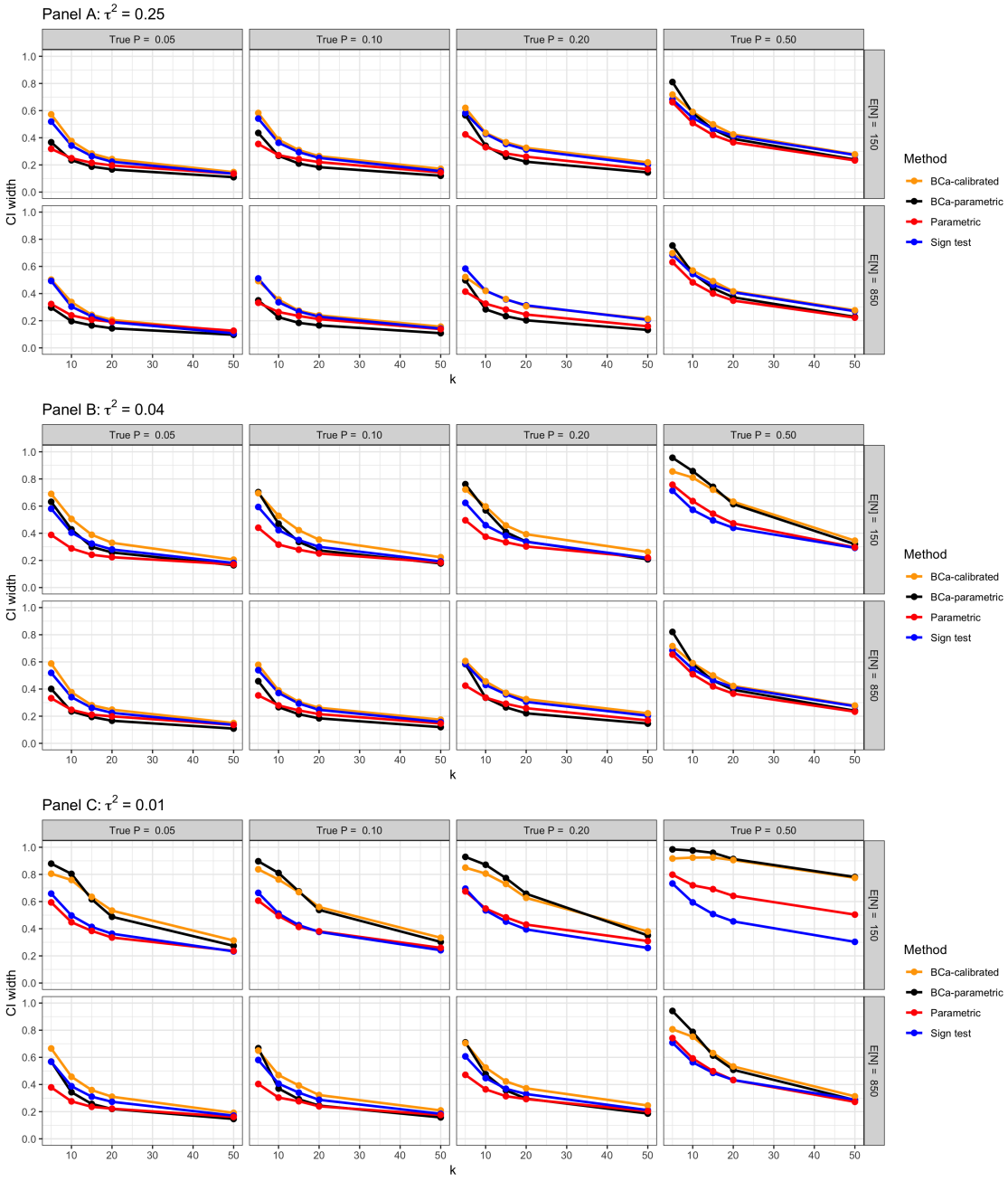
eFigure 6: *Width of 95% confidence intervals for exponential distribution*



eFigure 7: *Width of 95% confidence intervals for normal distribution*

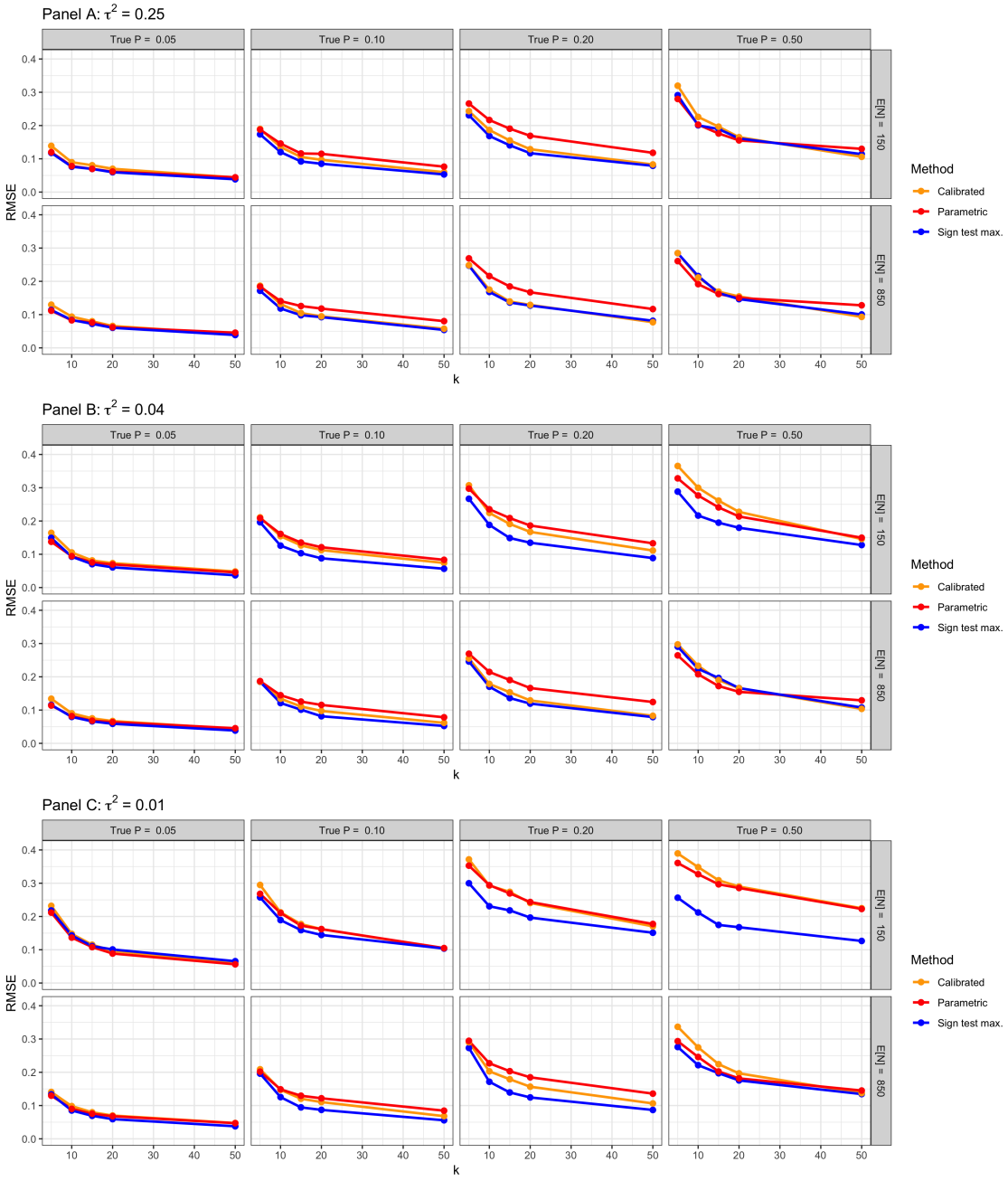


eFigure 8: *Width of 95% confidence intervals for scaled t distribution*

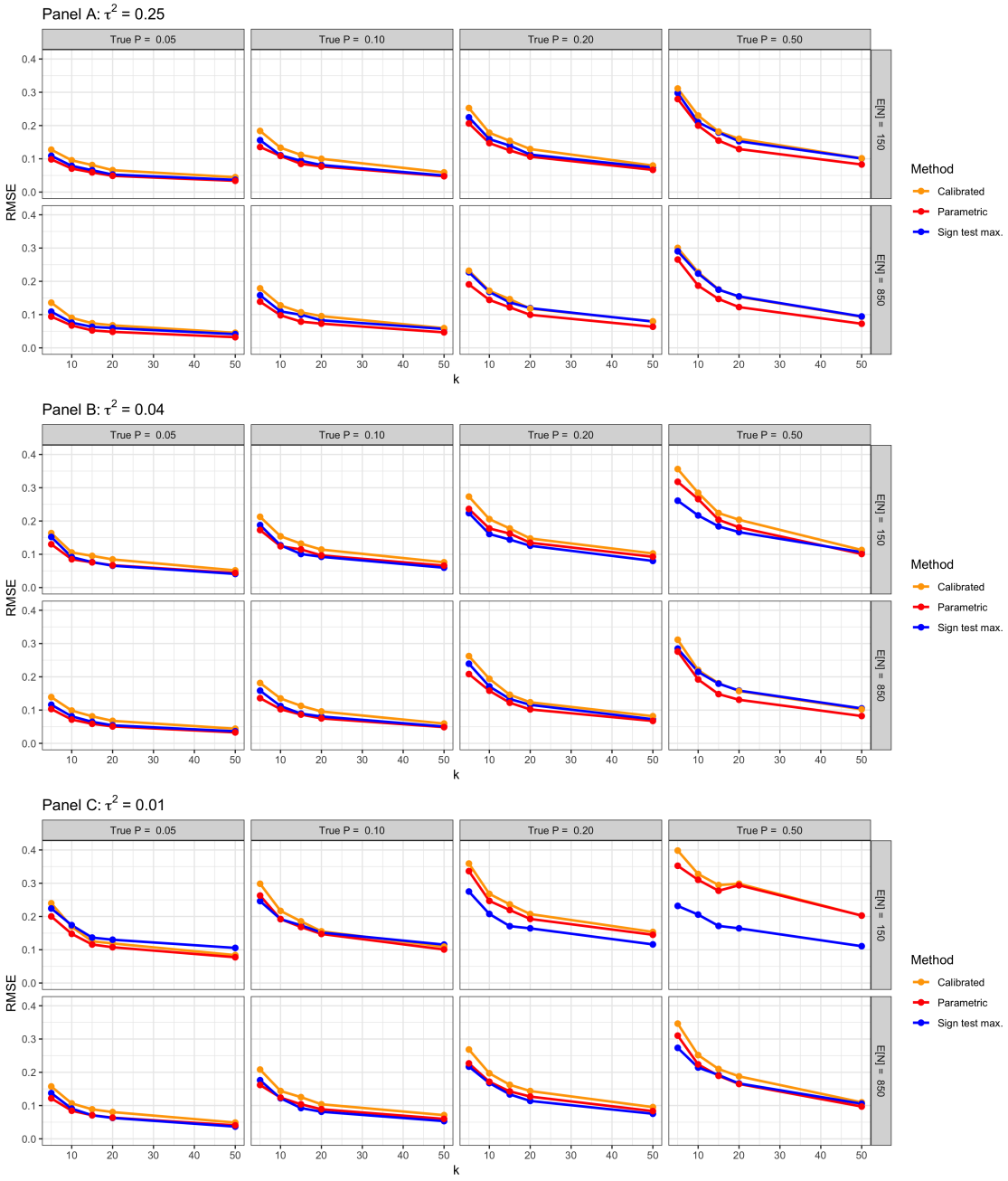


eFigure 9: *Width of 95% confidence intervals for uniform mixture distribution*

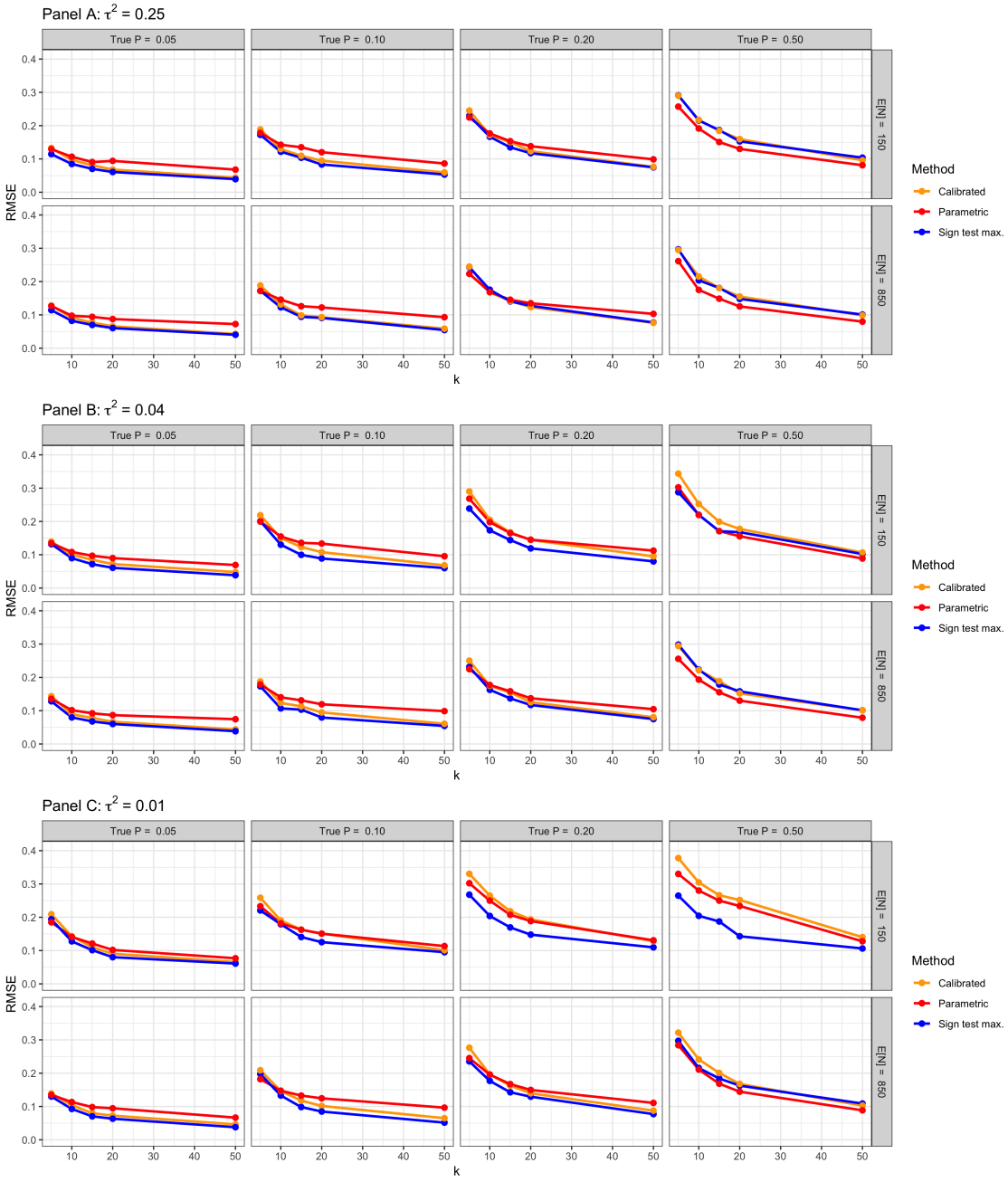
2.5. All RMSE results by distribution



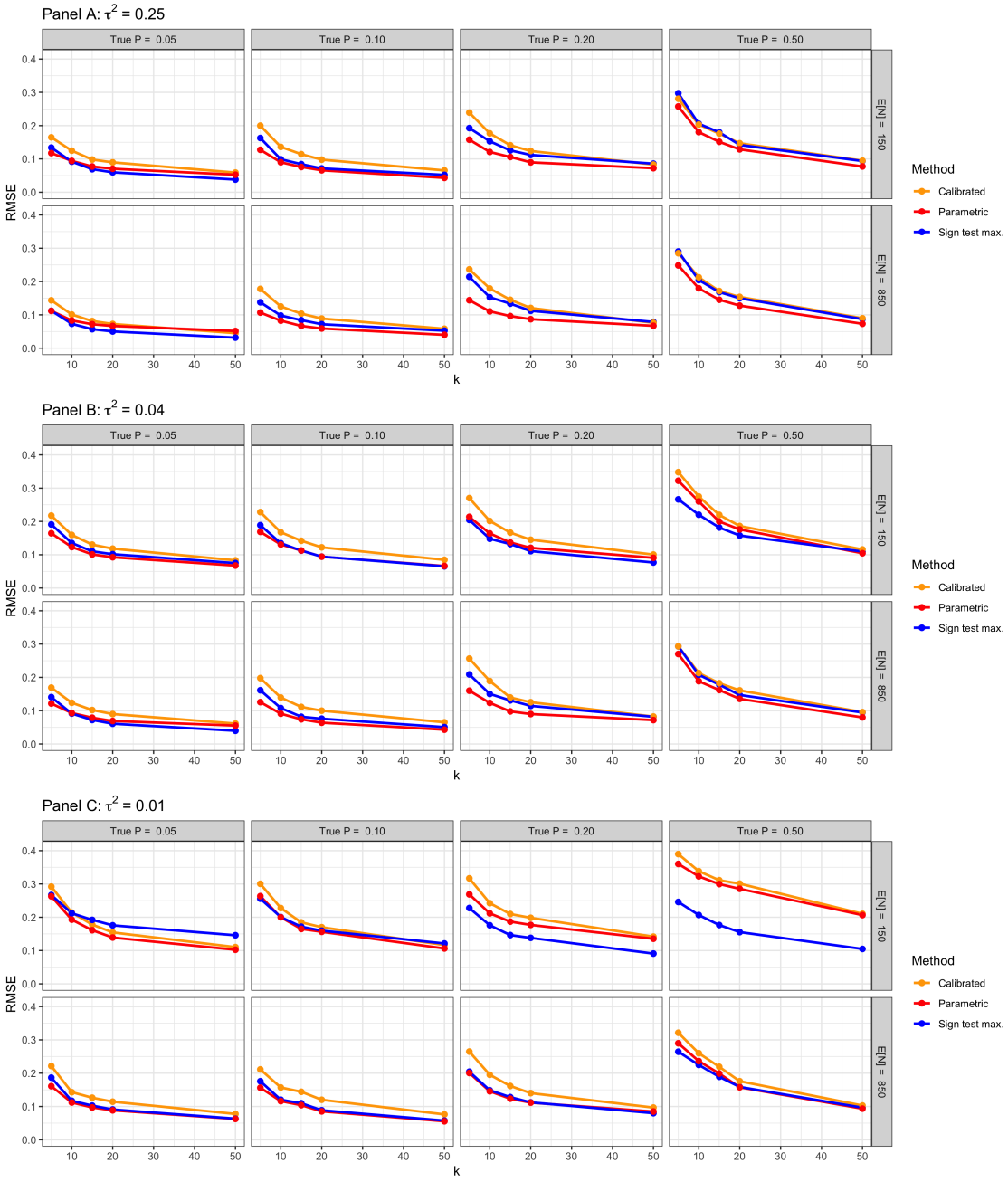
eFigure 10: *RMSE of point estimates for exponential distribution*



eFigure 11: *RMSE of point estimates for normal distribution*

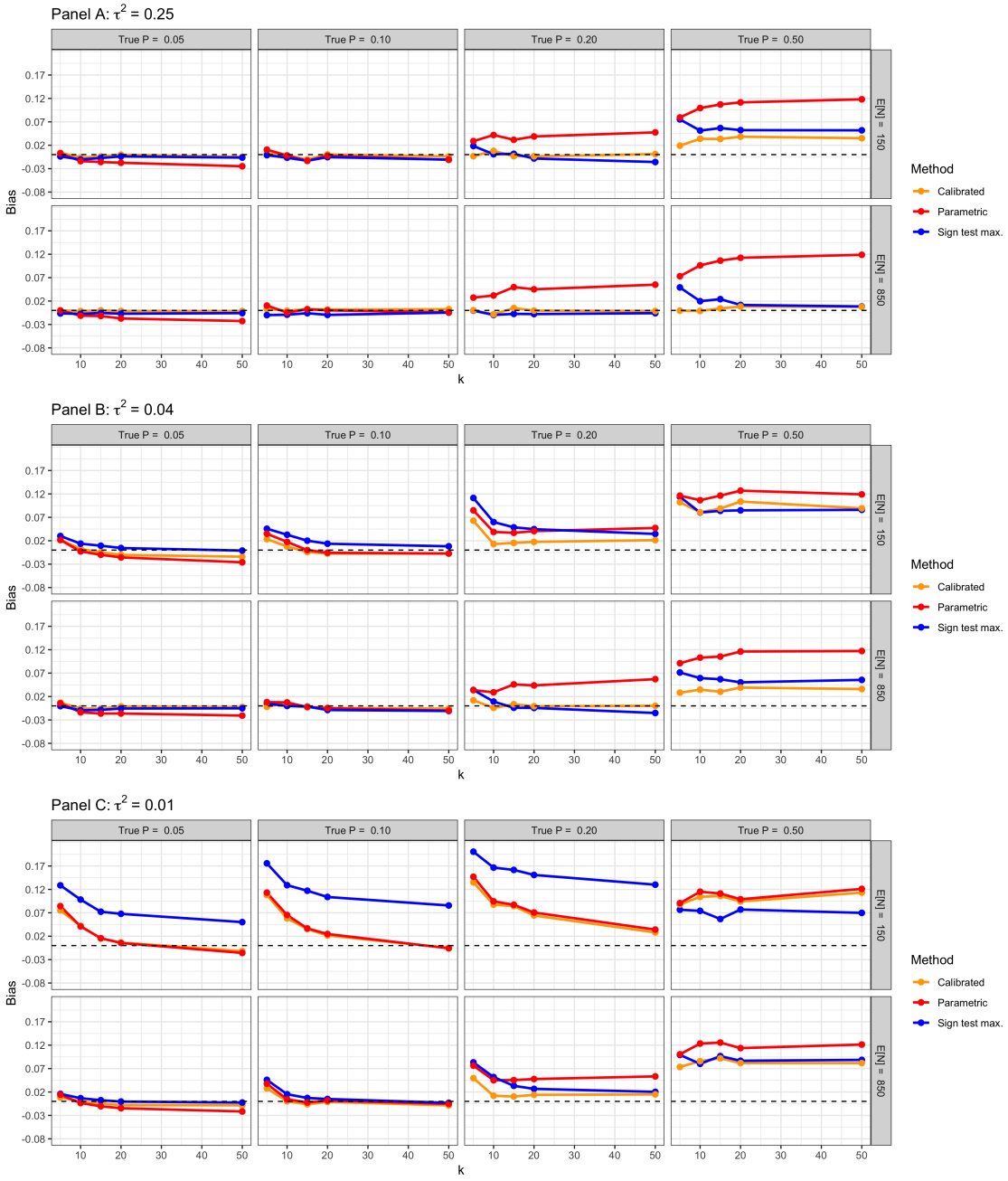


eFigure 12: *RMSE of point estimates for scaled t distribution*

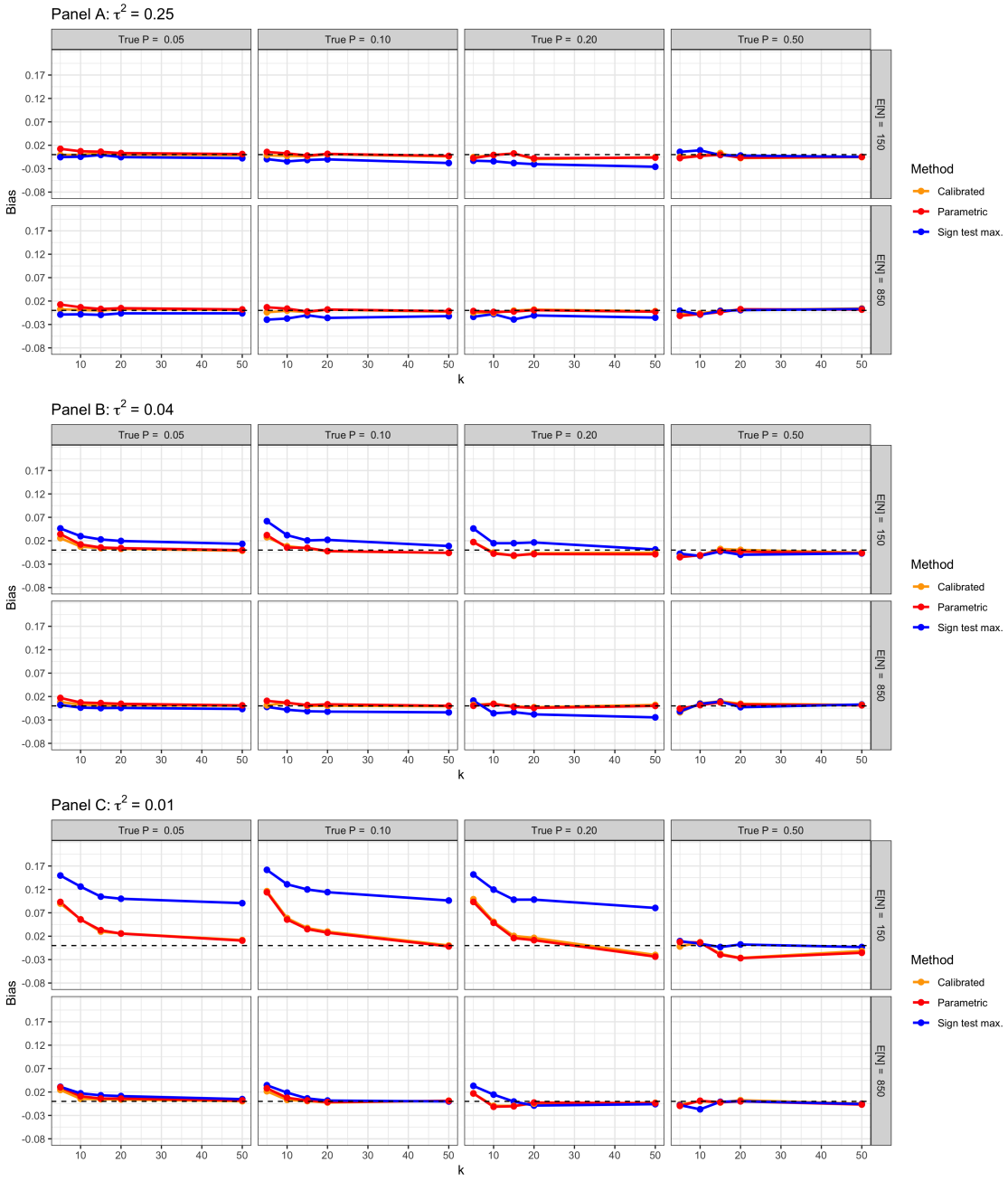


eFigure 13: *RMSE of point estimates for uniform mixture distribution*

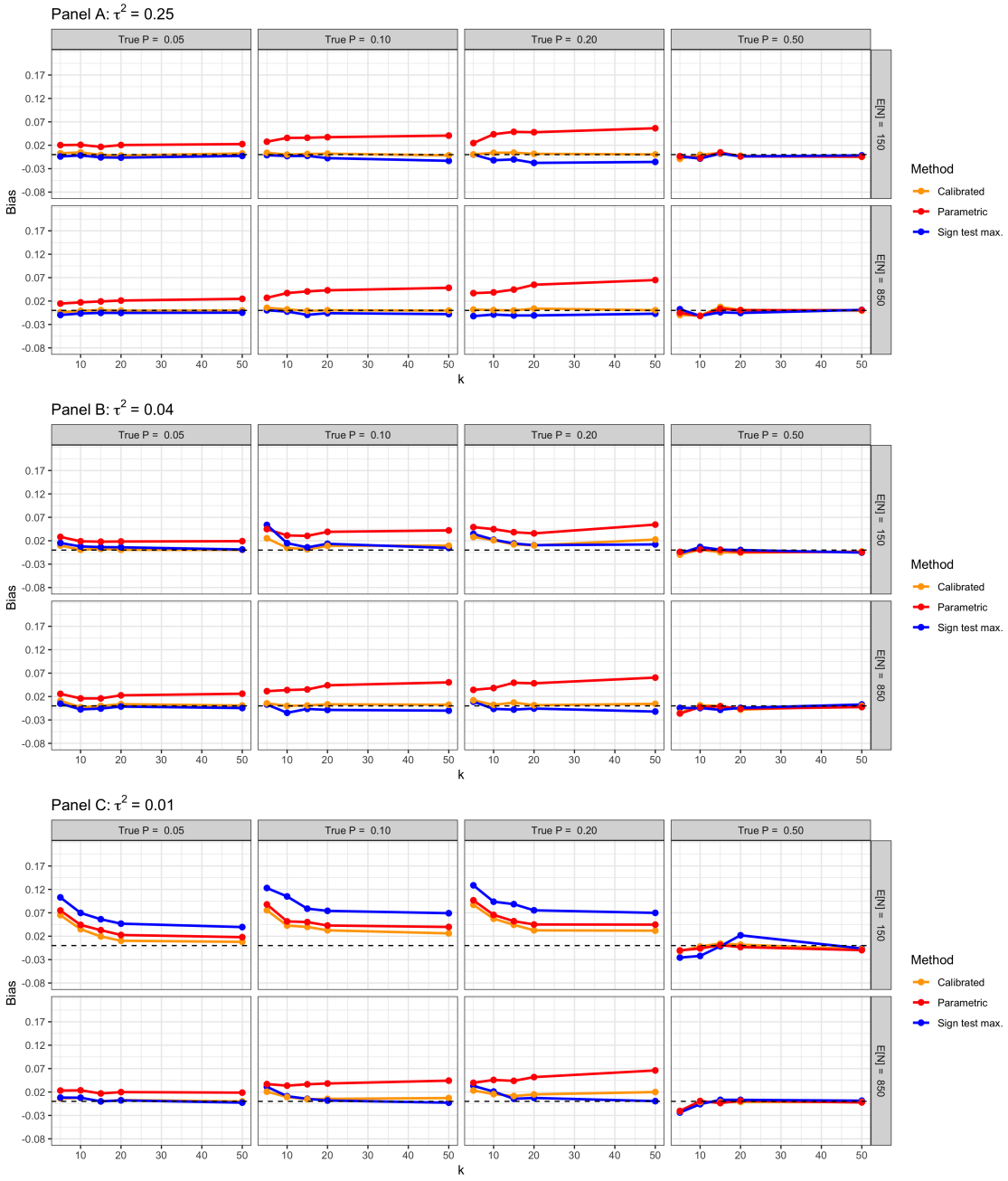
2.6. All bias results by distribution



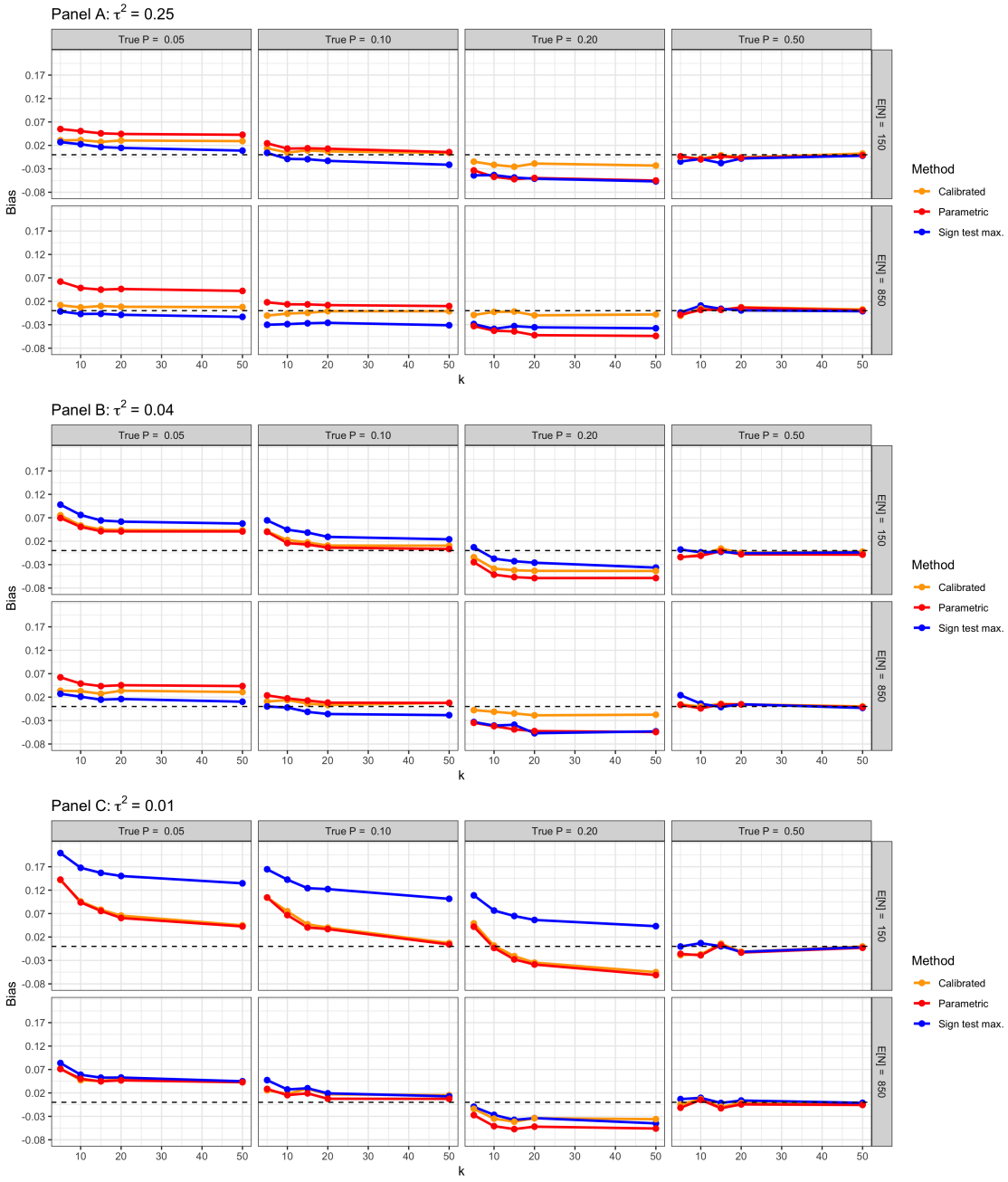
eFigure 14: *Bias of point estimates for exponential distribution*



eFigure 15: *Bias of point estimates for normal distribution*



eFigure 16: *Bias of point estimates for scaled t distribution*

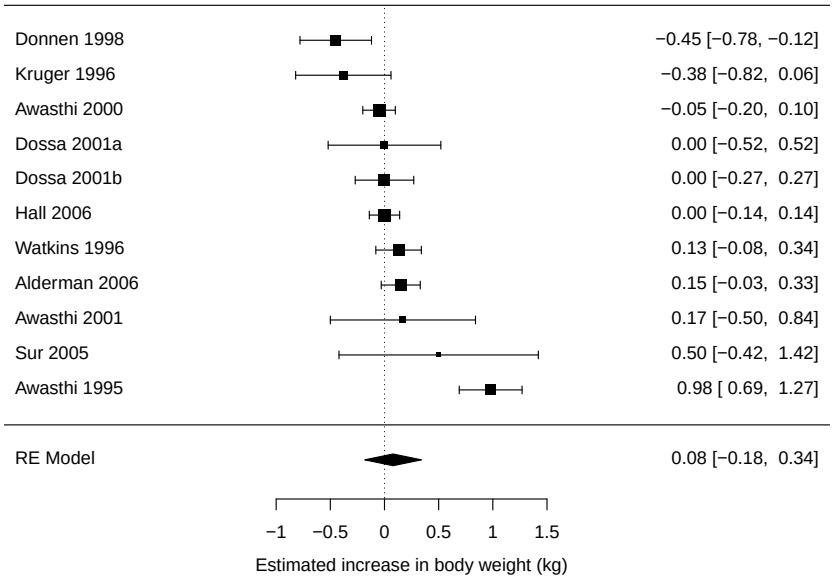


eFigure 17: *Bias of point estimates for uniform mixture distribution*

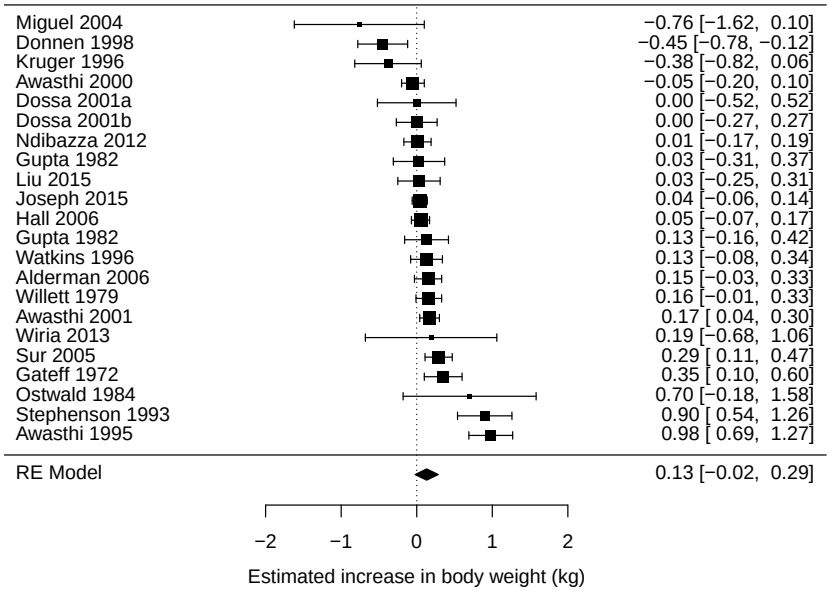
3. Applied example

We illustrate the methods using two meta-analyses of randomized controlled trials on the effect of mass deworming programs in developing countries on children's bodyweights. A Cochrane meta-analysis[9] of 11 studies reported "little to no effect" (estimated mean increase in bodyweight [kg]: 0.08; 95% CI: [-0.11, 0.27]), while an updated meta-analysis of 22 studies by other investigators[10] reported "significant" mean increases in weight (estimate: 0.13; 95% CI: [0.03, 0.24]). A largely unresolved controversy ensued[11]. We obtained data for both meta-analyses[10,9] from Croke et al.'s (2016)[10] Figures 1 and 2; the data are reproduced in Figures 18a and 18b. (We fit random-effects meta-analysis models using restricted maximum likelihood with standard errors adjusted via the Knapp-Hartung method in keeping with best practices[12,13]; hence, the confidence intervals reported in our analyses are slightly wider than those the original authors reported using the Dersimonian-Laird method[10], which are the estimates reported above.) Figure 19 shows estimated densities of the standardized point estimates, $(\widehat{\theta}_i - \widehat{\mu})/\sqrt{\widehat{\tau}^2 + \widehat{\sigma}_i^2}$, in each meta-analysis[14] and suggests some non-normality in both cases. Similarly, Shapiro-Wilk tests on the standardized point estimates yielded $p = 0.11$ and $p = 0.02$ for the Taylor-Robinson et al. (2015)[9] and Croke et al. (2016)[10] meta-analyses respectively, also suggesting some non-normality[15].

We therefore used the calibrated estimates to estimate the proportions of effects in each meta-analysis above and below several effect size thresholds, and we used the BCa-calibrated method for confidence intervals. For all thresholds, the meta-analyses in fact seemed to agree closely (Table 4). For example, both suggested that a majority of effects are above 0 ($\widehat{P}_{>0}$ = 0.73 with 95% CI: [0, 1] and 0.82 with 95% CI: [0.41, 0.91] respectively[9,10]); these point estimates suggest frequent beneficial effects of mass deworming, albeit possibly of very small size. The meta-analyses also both suggested that a sizable minority of effects are above 0.2 kg ($\widehat{P}_{>0.2}$ = 0.18 with 95% CI: [0, 0.55] and 0.23 with 95% CI: [0, 0.41]), though the wide confidence intervals indicated that there was considerable uncertainty. The analysis also indicated that in at least some settings, the programs may in fact *decrease* bodyweight by at least 0.2 kg on average ($\widehat{P}_{<-0.2}$ = 0.18 with 95% CI: [0, 0.45] and 0.14 with 95% CI: [0, 0.36]). Although the point estimates appear to agree closely, some of the confidence intervals for the first meta-analysis[9] are quite wide, indicating considerable uncertainty. Table 4 illustrates how these results compare to those obtained using parametric methods.

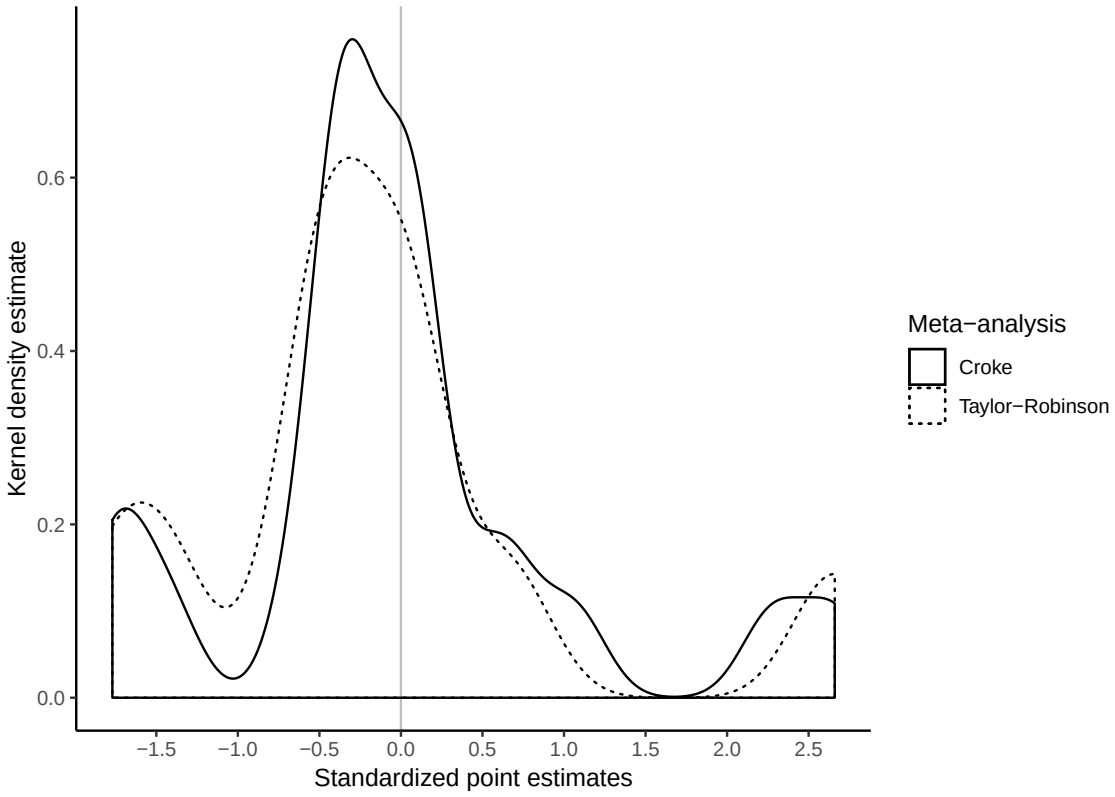


(a) *Taylor-Robinson et al. (2015)[9] meta-analysis*



(b) *Croke et al. (2016)[10] meta-analysis*

eFigure 18: *Forest plots for each meta-analysis of study-level point estimates (mean differences in bodyweight with vs. without mass deworming intervention) with 95% confidence intervals. Pooled point estimates were estimated via random-effects meta-analysis.*



eFigure 19: *Estimated densities of true population effects in each meta-analysis.*

	Taylor-Robinson (2015)		**Croke et al. (2016)**	
k	11		22	
$\widehat{\mu}$	0.08 [-0.18, 0.34]		0.13 [-0.02, 0.29]	
$\widehat{\tau}$	0.35 [0.00, 0.51]		0.28 [0.13, 0.38]	
Est. % of effects	**Parametric**	**Calibrated**	**Parametric**	**Calibrated**
> 0	59 [33, 85]	73 [0, 100]	68 [49, 87]	82 [41, 91]
> 0.1	48 [21, 74]	45 [0, 82]	54 [34, 75]	50 [14, 68]
> 0.2	36 [11, 62]	18 [0, 55]	40 [20, 61]	23 [0, 41]
> 0.5	11 [0, 58]	9 [0, 27]	10 [0, 35]	9 [0, 23]
< -0.1	30 [5, 55]	18 [0, 36]	21 [3, 38]	14 [0, 27]
< -0.2	21 [0, 44]	18 [0, 45]	12 [0, 29]	14 [0, 36]

eTable 4: *Number of studies (k), pooled point estimates ($\widehat{\mu}$), heterogeneity estimates of standard deviation of true population effects ($\widehat{\tau}$), and estimated percent of estimates above and below various thresholds ($100\% \times \widehat{P}_{>q}$ or $100\% \times \widehat{P}_{<q}$). "Parametric": point estimate was obtained parametrically and inference was obtained either via the delta method or by bootstrapping parametric estimates when the estimated percentage of effects was less than 15% or greater than 85%. "Calibrated": point estimate and inference were obtained using the calibrated estimates and BCa bootstrapping. Effect sizes are presented on the raw mean difference scale (kg of bodyweight). Brackets denote 95% confidence intervals.*

4. Software

The code below illustrates use of the function `prop_stronger` in the R package `MetaUtility` to estimate the proportion of effects above 0.5 in the Taylor-Robinson (2015) meta-analysis described above. The standard R documentation for the function provides details.

```
# see the Open Science Framework repository for the dataset "dt"
# and a few steps of data prep

library(MetaUtility)  # we ran version 2.0.0

# estimate proportion of effects above 0.5 kg
# using the recommended methods (i.e., calibrated estimates for
#  point estimate and BCa-calibrated method for CI)
prop_stronger(q = 0.5,  # threshold
              tail = "above",  # look at effects above the threshold
```

```
estimate.method = "calibrated",
ci.method = "calibrated",
dat = dt,  # dataset
yi.name = "yi",  # name of point estimate variable in dataset
vi.name = "vyi")  # name of study variance variable in dataset
```

References

[1] Rui Wang, Lu Tian, Tianxi Cai, and LJ Wei. Nonparametric inference procedure for percentiles of the random effects distribution in meta-analysis. *The Annals of Applied Statistics*, 4(1):520, 2010.

[2] Chia-Chun Wang and Wen-Chung Lee. A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*, 10(2):255–266, 2019.

[3] Maya B Mathur and Tyler J VanderWeele. Sensitivity analysis for unmeasured confounding in meta-analyses. *Journal of the American Statistical Association*, pages 1–20, 2019.

[4] Maya B Mathur and Tyler J VanderWeele. New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, 2018.

[5] Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

[6] Tyler VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, pages doi: 10.7326/M16–2607, 2017.

[7] Julian PT Higgins, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560, 2003.

[8] Jeffrey C Valentine, Therese D Pigott, and Hannah R Rothstein. How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2):215–247, 2010.

[9] David C Taylor-Robinson, Nicola Maayan, Karla Soares-Weiser, Sarah Donegan, and Paul Garner. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. *Cochrane Database of Systematic Reviews*, (7), 2015.

[10] Kevin Croke, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel. *Does mass deworming affect child nutrition? Meta-analysis, cost-effectiveness, and statistical power*. The World Bank, 2016.

[11] Muhammad Farhan Majid, Su Jin Kang, and Peter J Hotez. Resolving "worm wars": An extended comparison review of findings from key economics and epidemiological studies. *PLoS Neglected Tropical Diseases*, 13(3):e0006940, 2019.

[12] Joachim Hartung and Guido Knapp. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12):1771–1782, 2001.

[13] Joanna IntHout, John PA Ioannidis, and George F Borm. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14(1):1, 2014.

[14] Rebecca J Hardy and Simon G Thompson. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17(8):841–856, 1998.

[15] Samuel S Shapiro and RS Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.