Supplementary material: Simulation results

June 25, 2014

This appendix describes several simulation studies that characterize the performance of the c-REML effect size estimator. The first simulation compares the operating characteristics of the c-REML estimator and the HPS estimator under data-generating model MB1. The second and third simulations examine the operating characteristics of the c-REML estimator under models MB2 and MB4, respectively. For each simulation, we study (1) the bias and precision of the effect size estimators, (2) the bias of the associated variance estimators, and (3) the actual coverage rates of 95% confidence intervals.

In all three simulations, we used the `lme` function from the R package `nlme` (Pinheiro, Bates, DebRoy, & Sarkar, 2012, Version 3.1.111) to obtain REML estimates. This function uses a log-Cholesky parameterization for the random effects covariance matrix, which has an unrestricted parameter space (Pinheiro & Bates, 1996). We allowed the maximization routine to run for at most 50 iterations, and accepted the resulting values even if they had not converged to a maximum. We used the expected information matrix to evaluate the covariance of the REML variance component estimates $\mathbf{C}(\hat{\boldsymbol{\omega}})$ and the degrees of freedom ν.[1] Throughout, we display selected margins rather than results for every factor level. Complete simulation results, as well as the R code used to perform the simulations, are available in the accompanying R package `scdhlm`.

[1] The automatic output of the `nlme` package does not use the inverse expected information matrix for estimating the variance of the variance components, but rather uses a numerical approximation to the Hessian of the log-likelihood. We wrote a separate function to calculate the expected information matrix from supplied parameter estimates. The function is available in the accompanying R package `scdhlm`.

Table S1: Simulation design for model MB1

Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	8	-0.7	0.2	0.7
ρ	Within-case reliability	5	0.0	0.2	0.8
m	Number of cases	4	3	1	6
N	Measurement occasions	2	8	8	16

1 Model MB1

The first simulation study compared the performance of the c-REML estimator to that of the estimator proposed by HPS, using bias and mean-squared error as the criteria. Following the simulation studies reported by HPS, we used a multiple baseline design in which treatment assignment times are spread as evenly as possible across the range of measurement occasions while maintaining at least 3 measurement occasions within each phase.

Table 1 reports the design of the simulation study, which consists of an $8 \times 5 \times 4 \times 2$ factorial and follows the broad outlines of the simulations reported by HPS, but with fewer levels for certain parameters. Based on recent empirical research on levels of auto-correlation in single-case series (Shadish, Rindskopf, Hedges, & Sullivan, 2013; Shadish & Sullivan, 2011), the auto-correlation ϕ was varied between -0.7 and 0.7 (HPS used -0.9 to 0.9). The total variance was fixed to $\tau_0^2 + \sigma^2 = 1$ while the within-case reliability $\rho = \tau_0^2 / (\tau_0^2 + \sigma^2)$ was varied between 0.0 and 0.8. The number of cases and number of measurement occasions was limited to the smaller of the levels considered by HPS, because the bias of their estimator was negligible for larger sample sizes. The fixed effects were set to $\gamma_{00} = 0, \gamma_{10} = 1$, so that the effect size parameter δ_{AB} was equal to one. HPS observed that the bias of their estimator was proportional to the effect size parameter. On the assumption that the c-REML estimator behaves similarly, we interpret the simulated biases as proportions of the effect size parameter (e.g., bias of less than 2%).

For each combination of parameter levels, we generated 2×10^4 datasets. We then calculated two effect size estimates based on each simulated dataset: the HPS estimator

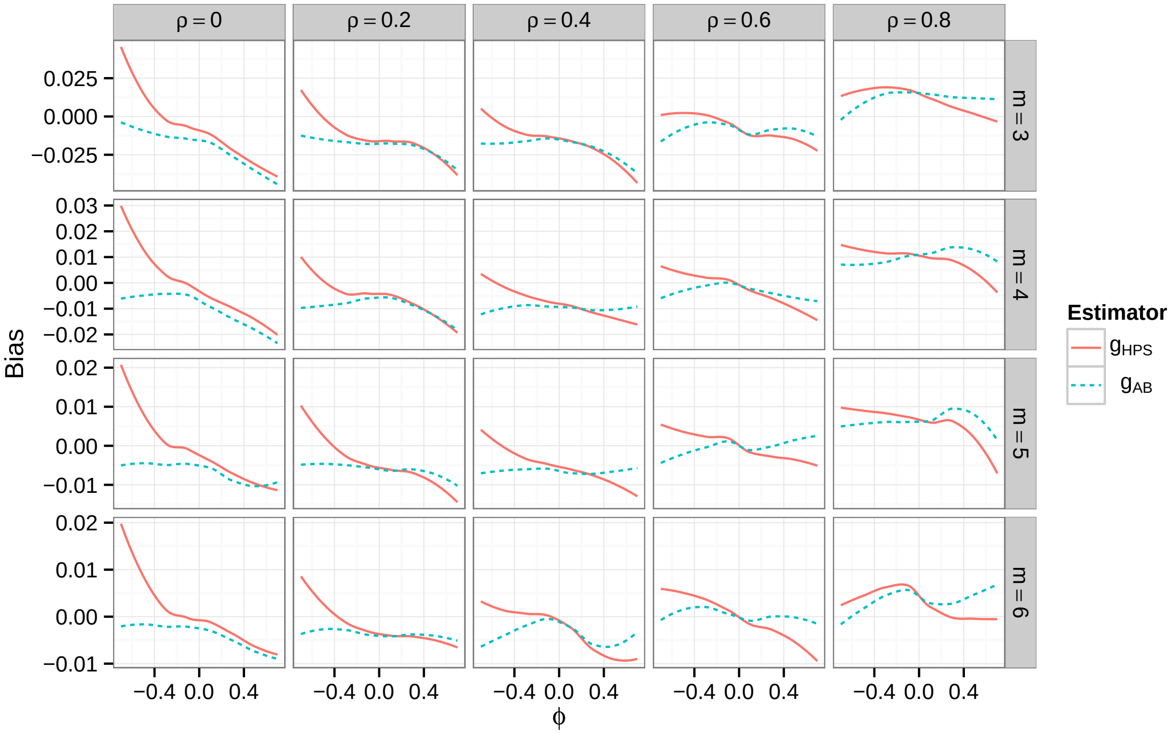

Figure S1: Bias of effect size estimators under MB1. Point-wise Monte Carlo s.e. < 0.005.

(denoted g_{HPS}) and the c-REML estimator (denoted g_{AB}). Both effect size estimators were calculated based on the same simulated data set, creating inter-correlation in their sampling distributions; consequently, differences between the estimators have very low Monte Carlo error.

Figure S1 plots the bias of the effect size estimators for the smaller series length of $N = 8$, across varying levels of the remaining parameters.[2] Overall, the bias of the c-REML estimator is quite small and comparable to the bias of the HPS estimator. At the smallest sample size considered ($m = 3, N = 8$), g_{AB} has a slightly more negative bias than g_{HPS} when ρ is low, but a comparable bias when ρ is larger; still, the bias of g_{AB} is never greater than 4.3% in absolute magnitude. For $m = 4$, the bias of g_{AB} is always less than 2.4%, while at the largest sample size considered ($m = 6$), the bias does not exceed 0.9%.

[2] The rows of the lattice correspond to different values of the number of cases m, while the columns of the lattice correspond to varying values of ρ; the x-axis of each panel corresponds to ϕ, and different colors and line-types correspond to each estimator. Note that the vertical axis of each graph differs across rows. Results for $N = 16$ are very similar.

Table S2: Average mean-squared error of effect size estimators under MB1

	$N = 8$		$N = 16$	
m	g_{HPS}	g_{AB}	g_{HPS}	g_{AB}
3	0.265	0.231	0.200	0.168
4	0.173	0.155	0.135	0.114
5	0.128	0.115	0.099	0.084
6	0.102	0.092	0.078	0.067

Given that both estimators have small biases, it is reasonable to also compare their precision. Table 1 reports the average mean squared error (MSE) of each estimator, where the average is taken over the levels of the nuisance parameters ϕ and ρ. On average and across sample sizes, g_{AB} has slightly better precision than the HPS estimator.

In addition to the effect size estimate itself, an estimate of its sampling variance is needed for meta-analysis. We assessed the performance of proposed variance estimators using relative bias; for an effect size estimator g with associated variance estimator V_g, the relative bias of the variance estimator is the ratio of the expected value of the variance estimator $\mathrm{E}(V_g)$ to the true variance of the effect size estimator $\mathrm{Var}(g)$. Relative biases close to one mean that the variance estimator is unbiased.

Figure S2 plots the relative bias of the c-REML and HPS variance estimators when $N = 8$, and is constructed in the same fashion as Figure S1.[3] From the figure, it can be seen that the HPS variance estimator is somewhat inaccurate for smaller values of the within-case reliability ρ, tending to under-estimate the true variance for positive values of the autocorrelation ϕ; this is true even at the larger sample size considered. In contrast, the variance estimator for g_{AB} provides more accurate estimates, with bias that depends less strongly on ρ and ϕ. Averaging across the levels of the nuisance parameters, the variance estimator for g_{AB} tends to over-estimate the true variance by 16% at the smallest sample size considered ($m = 3$ and $N = 8$). However, the average relative bias is close to unity for larger numbers of cases: for $m = 4$, the average relative bias is 1.03 when $N = 8$ and

[3] Results for $N = 16$ are very similar.

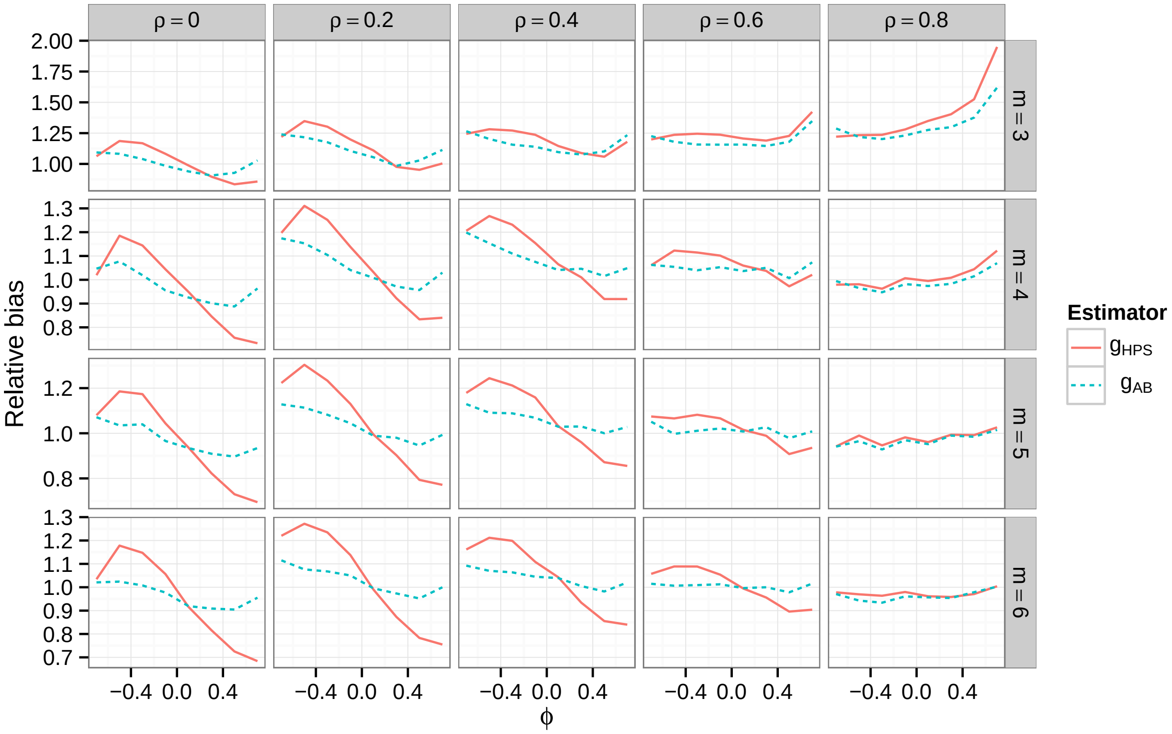

Figure S2: Relative bias of variance estimators under MB1. Point-wise Monte Carlo s.e. < 0.021.

0.98 when $N = 16$. Together with the small biases displayed by g_{AB}, these results suggest that the c-REML estimator is a reasonable alternative to the methods described by HPS for estimating effect sizes based on MB1.

Next, we examined the coverage rates of the two CIs described in the main text of the article. Figure S3 reports the coverage rates of 95% confidence intervals constructed using symmetric t critical values and using the non-central t critical values; each boxplot represents the distribution of coverage rates across the levels of the nuisance parameters. The symmetric CI tends to over-cover: with $m = 3$ the average coverage rate is 96.3%, decreasing to 95.6% when $m = 6$. The CI based on a non-central t approximation tends to have less than the stated coverage rate, with average rates ranging from 91.9% when $m = 3$ to 93.5% when $m = 6$. The coverage rate of the non-central CI is also more variable across the space of the nuisance parameters ϕ and ρ. For instance, when $m = 6$, the coverage of the symmetric CI

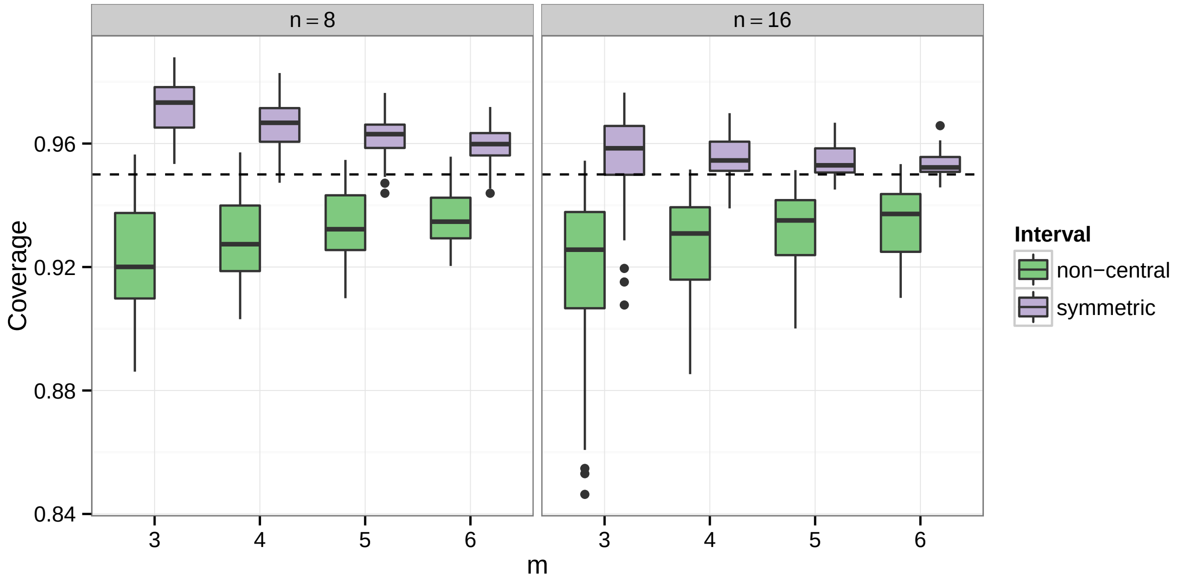

Figure S3: Range of coverage rates for nominal 95% confidence intervals under MB1. Monte Carlo error is negligible.

ranged from 94.4% to 97.2% across the levels of the nuisance parameters, while the coverage of the non-central CI was more variable, ranging from 91.0% to 95.6%.

Finally, some incidental results from this first simulation study shed further light the performance of the HPS variance estimation methods. HPS noted that the poor performance of their variance estimator may be due to the fact that it depends strongly on the values ϕ and ρ, which must typically be estimated from the data. For estimating these nuisance parameters, HPS used moment estimators that have poor sampling properties when phase lengths are short, regardless of the number of cases used. In comparison, the REML estimators of the same nuisance parameters are less biased and more precise. To illustrate this, Figure S4 plots the biases of the REML estimator and the estimator used by HPS as a function of the true parameter ϕ, specializing the results to $\rho = 0.6$. Both estimators are approximately unbiased when $\phi = 0$, but the moment estimator of ϕ has a large, negative bias that is approximately proportional to the true parameter value. Though the bias of the this estimator is mitigated by increasing the number of measurement occasions, it remains constant as the number of cases increases from $m = 3$ to $m = 6$. In light of this, we speculate

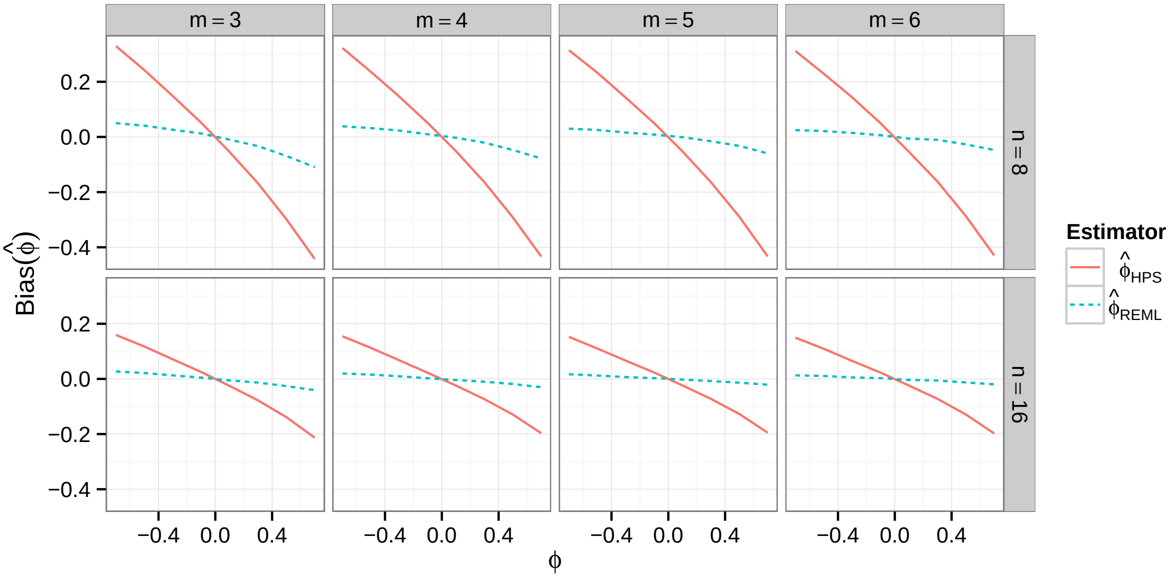

Figure S4: Bias of auto-correlation estimators under MB1 when $\rho = 0.6$. Point-wise Monte Carlo s.e. < 0.002.

that the performance of the variance estimator proposed by HPS may be improved by using different estimators of the nuisance parameters, such as the REML estimators.

2 Model MB2

The second simulation study examined the operating characteristics of the c-REML estimator under MB2. Table 2 reports the design of the simulation study, which consists of an $8 \times 5 \times 2 \times 4 \times 2$ factorial. Compared to MB1, MB2 has one further random effect, and thus two additional variance components: the variance of treatment effects τ_1^2 and the covariance of the treatment effects and baseline levels τ_{10}. In the previous simulation study, it was possible to explore the parameter space fairly thoroughly, but the greater number of parameters prohibits as comprehensive a simulation for MB2. We limited the simulation design in several ways. First, we parameterized the between-case variance in treatment effects as a proportion of the between-case variation in baseline levels; letting $\lambda_1 = \tau_1^2/\tau_0^2$, we set $\lambda_1 = 0.1$ or $\lambda_1 = 0.5$ to represent moderate and high levels of treatment effect heterogeneity,

Table S3: Simulation design for model MB2

Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	8	-0.7	0.2	0.7
ρ	Within-case reliability	5	0.0	0.2	0.8
λ_1	Ratio of variance components	2	0.1	0.4	0.5
m	Number of cases	4	3	1	6
N	Measurement occasions	2	8	8	16

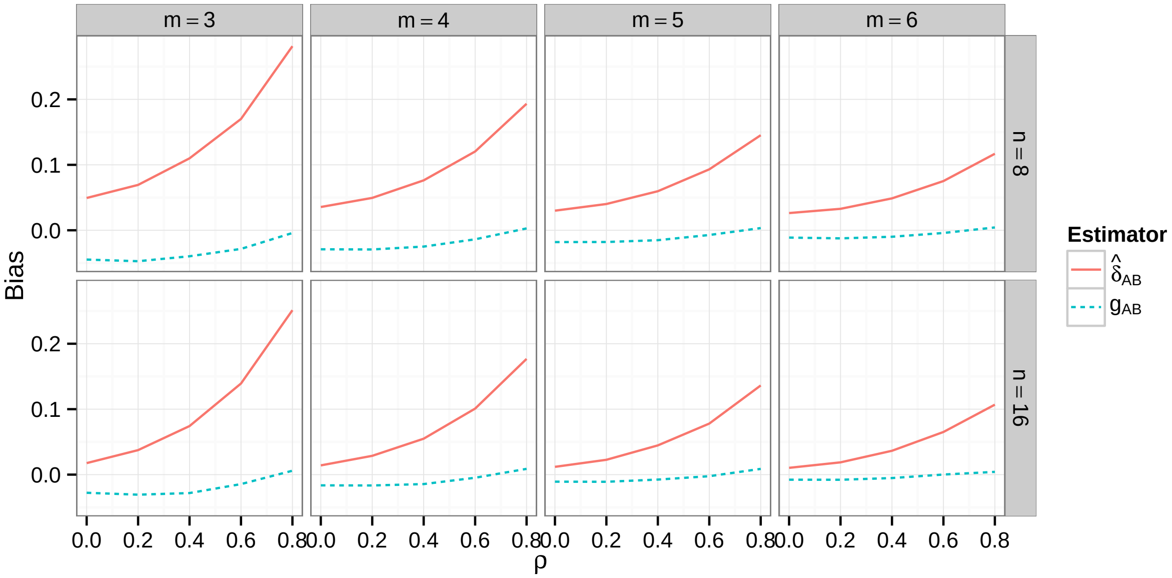

Figure S5: Bias of effect size estimators under MB2, averaged over levels of ϕ and λ_1. Pointwise Monte Carlo s.e. < 0.001.

respectively. Next, we set $\tau_{10} = 0$ because pre-testing indicated that the correlation between random effects had little influence on the bias of the effect size estimates. As previously, we set $\gamma_{00} = 0$, $\gamma_{10} = 1$, and $\tau_0^2 + \sigma^2 = 1$ so that the true effect size parameter $\delta_{AB} = 1$. For each combination of parameter levels, we generated 2×10^4 datasets and calculated the c-REML estimator, associated variance estimator, and confidence intervals for each dataset.

Figure S5 plots the bias of the c-REML estimator (g_{AB}) for varying numbers of cases m, measurement occasions N, and within-case reliability ρ, averaging over the values of the other parameters (ϕ and λ_1). Lacking other effect size estimators comparable to the HPS estimator for MB1, we used the unadjusted REML estimator ($\hat{\delta}_{AB}$) as a point of comparison;

Table S4: Average and maximum mean-squared error of g_{AB} under MB2

	$N = 8$		$N = 16$	
m	mean	max	mean	max
3	0.290	0.664	0.221	0.545
4	0.198	0.411	0.153	0.356
5	0.149	0.289	0.116	0.259
6	0.120	0.228	0.092	0.200

Figure S5 plots its bias as well. Even at the smallest sample size considered, the bias of g_{AB} is fairly small. When $m = 3$ and $N = 8$, the absolute bias is less than 7.3% across all combinations of parameters considered; for $m = 4$, the bias is always less than 4.9%; for $m \geq 5$, the bias is always less than 2.9%. The bias of g_{AB} is also both smaller and less variable than that of $\hat{\delta}_{AB}$, which increases substantially with ρ. Under this data-generating model, the c-REML estimator appears to have biases small enough to warrant use in meta-analysis, particularly for studies with at least four cases.

Alhough the c-REML estimator has only small biases, it is nonetheless imprecise when the sample of cases is small. To characterize its variability, Table 2 reports the average and maximum MSE of the estimator across the levels of the nuisance parameters. The average MSE is between 26% and 38% higher than the corresponding MSE under Model MB1. As a point of comparison, the average MSE when $m = 4$ and $N = 8$ is comparable to the variance of the Hedges' g effect size estimate from a balanced two-sample experiment with 25 participants; the average MSE when $m = 6$ and $N = 8$ is comparable to that from an experiment with 39 participants. The MSE of g_{AB} varies depending on the level of auto-correlation and within-case reliability, and is generally increasing in ϕ, ρ, and λ_1.

Next, Figure S6 displays the relative bias of the c-REML variance estimator versus the auto-correlation ϕ, for varying numbers of cases and measurement occasions. Across the parameter space and across levels of m, and N, the variance of g_{AB} tends to be over-estimated. The over-estimation is substantial for small samples of cases: for example, the average relative bias is 1.43 when $m = 3$ and $N = 16$. Even at the largest sample size considered, the

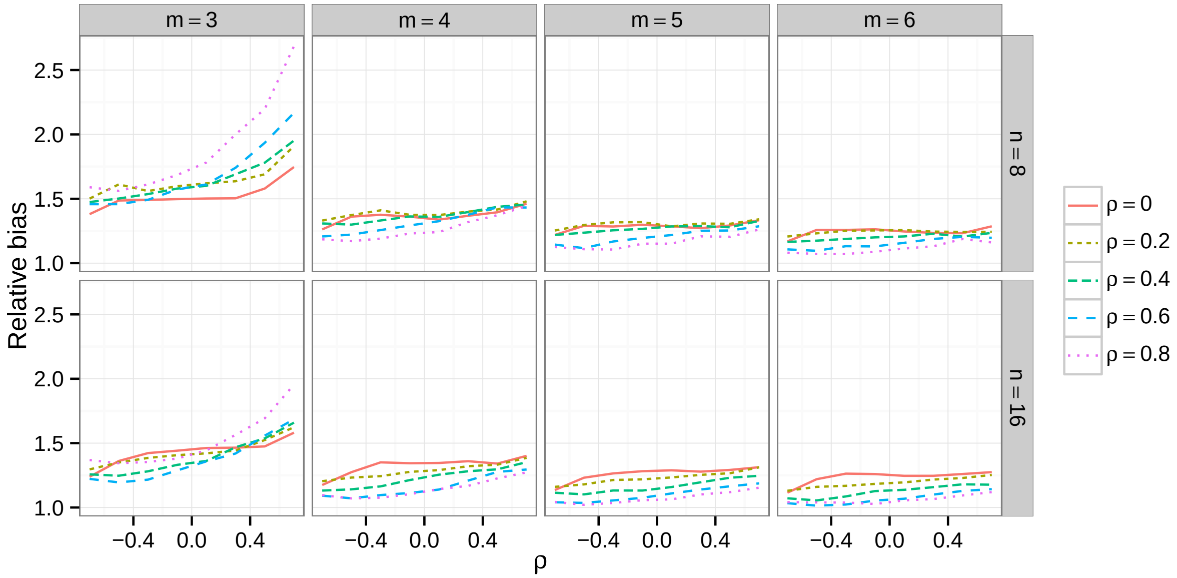

Figure S6: Relative bias of c-REML variance estimator under MB2, averaged over levels of λ_1. Separate lines are plotted for different levels of ho to illustrate how the relative bias depends on the within-case reliability. Point-wise Monte Carlo s.e. < 0.022.

variance is moderately over-estimated: the average relative bias is 1.14 when $m = 6$ and $N = 16$. The bias in the approximate variance estimator may come from several sources, including from approximating the distribution of $\hat{\delta}_{AB}$ by a t-distribution, from approximating the variance of the REML estimates using the inverse expected information matrix, and from using imprecise parameter estimates in calculating the information matrix.

Finally, Figure S7 reports the coverage rates of 95% confidence intervals constructed using symmetric t critical values and using the non-central t critical values. The results are broadly similar to the previous simulation. The symmetric CI tends to over-cover: with $m = 3$ the average coverage rate is 97.8% (with a minimum of 95.1% and a maximum of 99.5%), decreasing to 96.4% when $m = 6$ (with a minimum of 93.9% and a maximum of 97.6%). The CI based on a non-central t approximation tends to have less than the stated coverage rate, with average rates ranging from 93.3% when $m = 3$ (minimum: 88.5%, maximum 96.4%) to 94.2% when $m = 6$ (minimum: 92.0%, maximum 96.3%). The coverage rate of the non-central CI is also more variable across the space of the nuisance parameters.

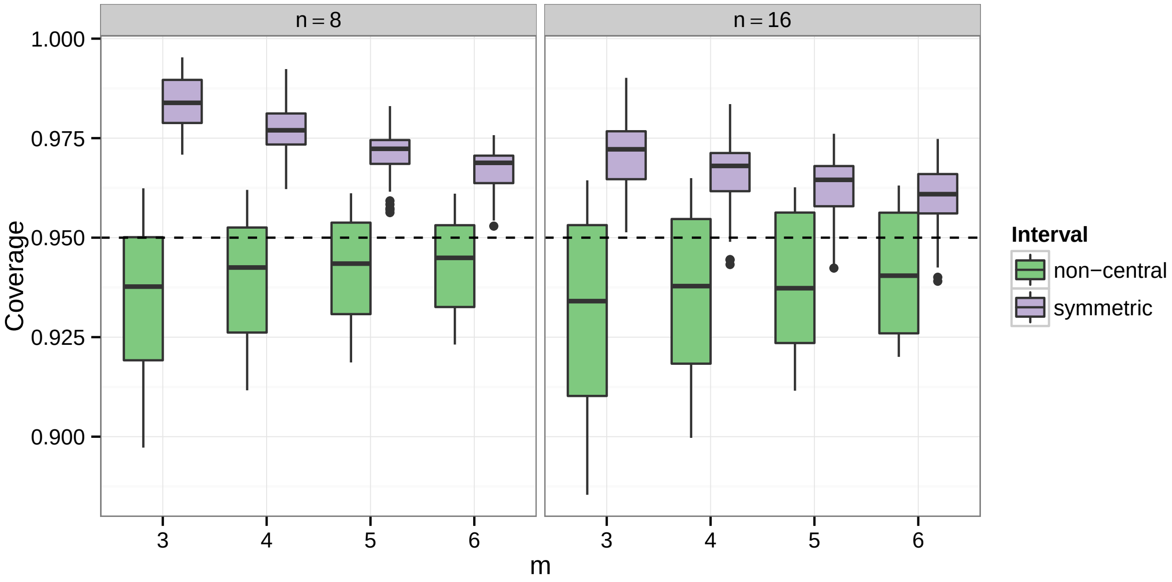

Figure S7: Range of coverage rates of 95% confidence intervals under MB2. Monte Carlo error is negligible.

3 Model MB4

The third simulation study examined the operating characteristics of the c-REML estimator under MB4, which allows for baseline trends that vary across cases as well as a change in trend due to treatment (constant across cases). Compared to MB1, MB4 has two additional parameters in the mean specification (the baseline trend γ_{20} and the trend-by-treatment interaction γ_{30}) and two additional variance parameters (the variance of the baseline slopes τ_2^2 and the covariance of the baseline slopes and levels τ_{20}).

The effect size parameter in MB4 depends on the analyst's choice of times A and B, which characterize a hypothetical between-case randomized experiment by the point of treatment introduction and the point of outcome measurement, respectively. For purposes of simulation, we took $A = N/2$, and $B = 3N/4$, so that the effect size represents the change due to treatment after $N/4$ measurement occasions, which is a reasonably short time relative to the length of the study. Because MB4 includes linear baseline time trends that vary across cases, the choice of centering point for the time trend affects the interpretation

Table S5: Simulation design for model MB4

Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	8	-0.7	0.2	0.7
ρ	Within-case reliability	5	0.0	0.2	0.8
λ_2	Ratio of variance components	2	0.1	0.4	0.5
m	Number of cases	6	3	1/3	12
N	Measurement occasions	2	8	8	16

of the variance components. We centered time at point $B = 3N/4$, so that τ_0^2 represents the between-case variation in the level of the outcomes at time B and τ_{20} is the covariance between case-specific baseline slopes and baseline outcome levels at time B.

Table 3 summarizes the design of the third simulation study, which parallels that for MB2. The main difference is that we extended the maximum number of cases considered, using $m = 3, 4, 5, 6, 9, 12$. To moderate dimensionality, we limited the parameter space in several ways. First, we parameterized the between-case variance in baseline slopes as a proportion of the between-case variation in baseline levels; letting $\lambda_2 = \tau_2^2/\tau_0^2$, we set $\lambda_2 = 0.1$ or $\lambda_2 = 0.5$ and used $\tau_{20} = 0$ throughout. Next, we did not vary the fixed effects, instead setting the average baseline outcome level $\gamma_{00} = 0$, the (fixed) change in the level due to treatment $\gamma_{10} = 1$, the average baseline slope $\gamma_{20} = 0$, and the (fixed) increase in slope due to treatment $\gamma_{30} = 0$. Finally, we set $\tau_0^2 + \sigma^2 = 1$ so that the true effect size parameter is $\delta_{AB} = 1$. For each combination of parameter levels, we generated 2×10^4 datasets and calculated the c-REML estimator, variance estimator, and confidence intervals for each dataset. The presentation of results parallels that of the previous simulation study.

Figure S8 plots the bias under MB4 of the c-REML estimator (g_{AB}) as well as the unadjusted estimator ($\hat{\delta}_{AB}$). As under MB2, the bias of g_{AB} is surprisingly small, even at the smallest sample sizes considered. When $m = 3$ and $N = 8$, the absolute bias is less than 5.8% across all combinations of parameters considered, though it is as large as 12.8% when $n = 16$. For $m = 4$, the bias is never more than 2.7%, and the average bias across the parameter space is only -0.8%. Also as under MB2, g_{AB} is substantially less biased than

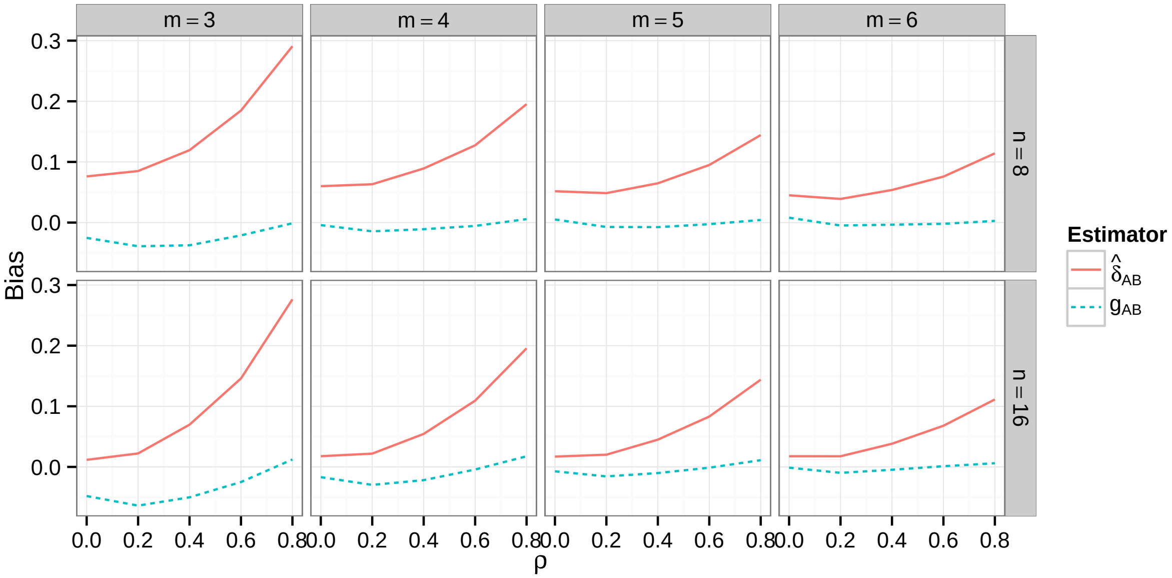

Figure S8: Bias of effect size estimators under MB4, averaged over levels of ϕ and λ_2. Pointwise Monte Carlo s.e. < 0.002.

$\hat{\delta}_{AB}$, which had a bias that increases with ρ. As previously, the bias is small enough that the point estimator g_{AB} should be considered suitable for use in meta-analysis, particularly for studies with $m \geq 4$.

Though the c-REML estimator has only small biases, it is nonetheless quite imprecise when the sample of cases is small. To characterize its variability, Table 3 reports the average and maximum MSE of the estimator across the levels of the nuisance parameters. The estimator has very high MSE for some combinations of parameters, and is generally increasing in ϕ, ρ, and λ_2. The average MSE is between 122% and 194% higher than the corresponding MSE under Model MB1. As a point of comparison, the average MSE when $m = 6$ and $N = 8$ is comparable to the variance of Hedges' g from a balanced, two-sample experiment with 20 participants. The high MSE is a consequence of the need to estimate a scale parameter (the denominator of the effect size) that includes between-case variance components, based on a sample of only a few cases.

Next, Figure S9 displays the relative bias of the variance estimator versus the autocorrelation ϕ, for varying numbers of cases and measurement occasions; its construction

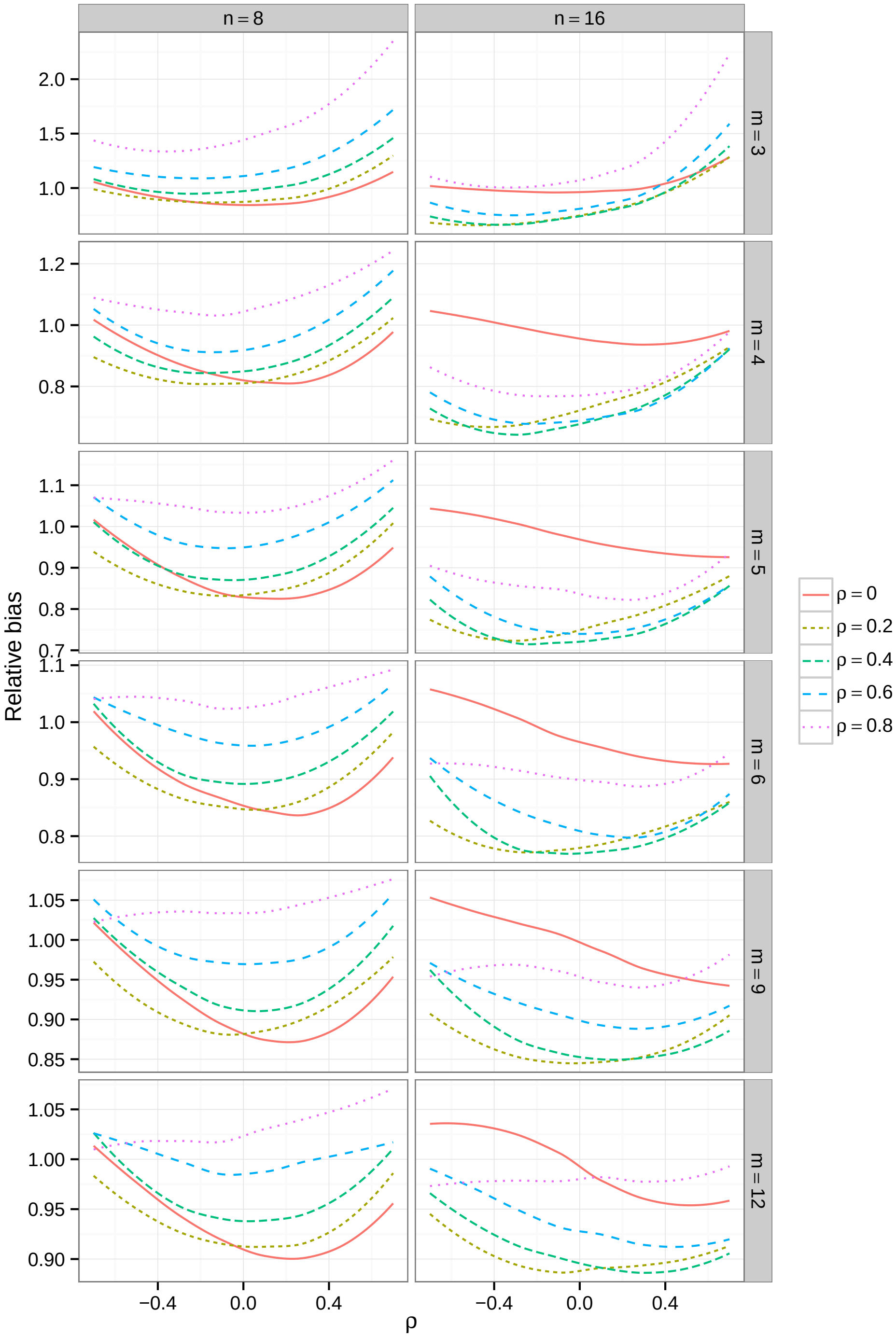

Figure S9: Relative bias of c-REML variance estimator under MB4, averaged over levels of λ_2. Note the scale of the vertical axis differs by lattice row. Point-wise Monte Carlo s.e. < 0.018.

Table S6: Average and maximum mean-squared error of g_{AB} under MB4

	$N = 8$		$N = 16$	
m	mean	max	mean	max
3	0.596	0.862	0.373	0.624
4	0.444	0.685	0.256	0.427
5	0.339	0.546	0.188	0.290
6	0.255	0.421	0.150	0.226
9	0.160	0.271	0.090	0.139
12	0.116	0.199	0.066	0.104

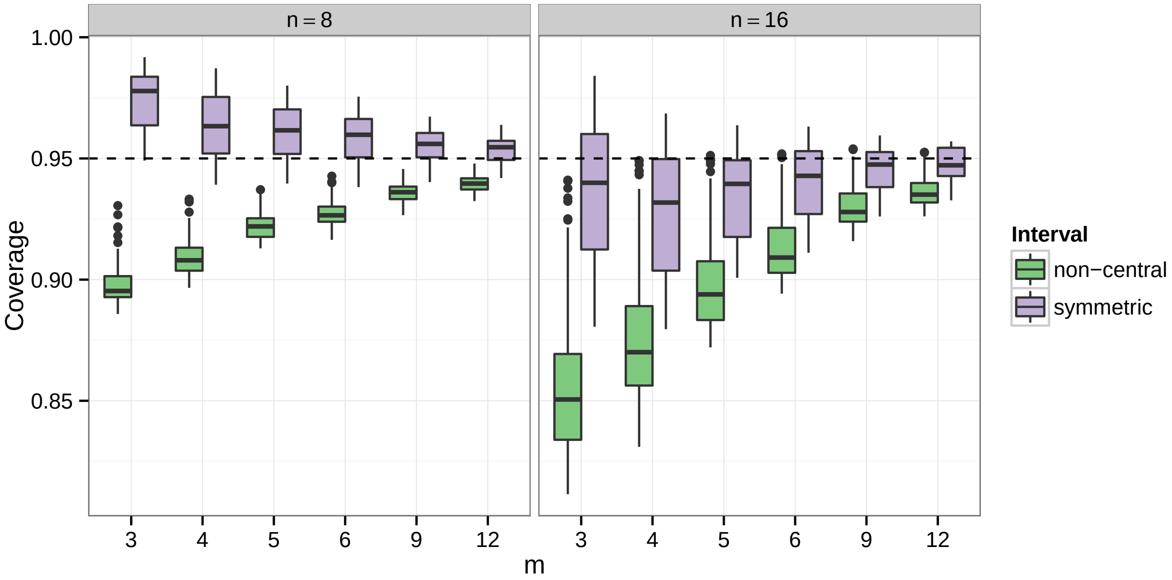

Figure S10: Range of coverage rates of 95% confidence intervals under MB4. Monte Carlo error is negligible.

parallels Figure S6, except that $m = 9$ and $m = 12$ are also included. Unlike in MB1 and MB2, here the variance approximation tended to have a downward bias, except for when $m = 3$ and $N = 8$. The under-estimation is more pronounced for longer series length. The bias of the variance estimator only becomes tolerably small when the number of cases is relatively large: even when $m = 6$, the average relative bias is 0.96 for $N = 8$ but 0.87 for $N = 16$. The variance approximation in this model depends on a multiple $n/4$ of the treatment-by-trend interaction γ_{30}, and so may be particularly sensitive to under-statement of the variance of fixed effects.

Finally, Figure S10 reports the average coverage rates of 95% confidence intervals constructed using symmetric t critical values and using the non-central t critical values. Unlike in previous simulations, the symmetric confidence interval does not always over-cover; while the average coverage rate is greater than 95% for the shorter series length of $N = 8$, it is less than 95% for the longer series length of $N = 16$. The coverage rate approaches nominal levels only when the number of cases is relatively large. For example, when $m = 6$, the coverage rate of the symmetric CI ranged from 91.1% to 97.5%; when $m = 9$, the coverage rate of the symmetric CI ranged from 92.6% to 96.7%. The CI based on a non-central t approximation tends to have less than nominal coverage. The coverage discrepancy is substantial when the number of cases is small or moderate, ranging from 87.9% when $m = 3$ (minimum: 81.1%, maximum 94.1%) to 92.1% when $m = 6$ (minimum: 89.4%, maximum 95.2%) and 93.3% when $m = 9$ (minimum: 91.6%, maximum 95.4%).

References

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, *6*, 289–296.

Pinheiro, J. C., Bates, D. M., DebRoy, S., & Sarkar, D. (2012). *nlme: Linear and Nonlinear Mixed Effects Models.* Retrieved from `http://cran.r-project.org/package=nlme`

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*(3), 813–821. doi: 10.3758/s13428-012-0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. doi: 10.3758/s13428-011-0111-y