DOI: 10.1037/met0000136

Multilevel Modeling of Single-Case Data:
A Comparison of Maximum Likelihood and Bayesian Estimation

Mariola Moeyaert

KU Leuven, University of Leuven, Belgium

David Rindskopf

City University of New York, New York

Patrick Onghena

KU Leuven, University of Leuven, Belgium

Wim Van den Noortgate

KU Leuven, University of Leuven & iMinds-itec, Belgium

Author Note

This research is funded by the Institute of Educational Sciences (Grants R305D110024 & R305D150007) and the Research Foundation –Flanders (FWO). The opinions expressed are those of the authors and do not represent the views of the Institute of Educational Sciences or the FWO.

Correspondence concerning this article can be addressed to Mariola Moeyaert, Department of Educational Psychology and Methodology, State University of New York, 1400 Washington Ave., Albany, NY 12222, USA.

E-mail mmoeyaert@albany.edu

Abstract

The focus of this paper is to describe Bayesian estimation, including construction of prior distributions, and to compare parameter recovery under the Bayesian framework (using weakly informative priors) and the maximum likelihood (ML) framework in the context of multilevel modeling of single-case experimental data. Bayesian estimation results were found similar to ML estimation results in terms of the treatment effect estimates, regardless of the functional form and degree of information included in the prior specification in the Bayesian framework. In terms of the variance component estimates, both the ML and Bayesian estimation procedures result in biased and less precise variance estimates when the number of participants is small (i.e., 3). By increasing the number of participants to 5 or 7, the relative bias is close to 5% and more precise estimates are obtained for all approaches, except for the inverse-Wishart prior using the identity matrix. When a more informative prior was added, more precise estimates for the fixed effects and random effects were obtained, even when only three participants were included.

Keywords: Bayesian statistics, maximum likelihood, weakly informative prior, single-case designs, two-level modeling

Multilevel Modeling of Single-Case Data: A Comparison of Maximum Likelihood and Bayesian Estimation

Over the past decades, single-case experimental design (SCED) studies have made significant contributions to educational policy and practice (Kratochwill et al., 2010) as they provide scientifically sound evaluations of treatment effect estimates (Kratochwill & Levin, 2010). In an SCED study, one entity or a small group of entities (i.e., subjects, participants, or experimental units) are the focus of interest and each entity is measured repeatedly during at least one baseline condition and one treatment condition. The main focus of SCEDs lies in assessing whether there is a causal relation between the introduction of a treatment and the change in a dependent variable (Levin, O'Donnell, & Kratochwill, 2003; Onghena, 2005). The most popular type of SCED study is the multiple-baseline design (MBD) across participants (Shadish & Sullivan, 2011, see Figure 1). In an MBD, an AB phase design (with one baseline phase, A, and one treatment phase, B) is implemented simultaneously to multiple participants (Barlow, Nock & Hersen, 2009; Ferron & Scot, 2005; Onghena, 2005). An inherent characteristic of MBDs is that the treatment is introduced sequentially across the participants. This entails the advantage that researchers can more easily disentangle treatment effects from external events, such as the illness of a teacher or the presence of a foreign observer (Baer, Wolf, & Risley, 1968; Barlow et al., 2009; Kinugasa, Cerin, & Hooper, 2004; Koehler, & Levin, 2000; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a).

--- INSERT FIGURE 1 ABOUT HERE ---

Combining research findings across participants within an SCED study can provide strong quantitative evidence for the overall effectiveness of treatments across participants (Shadish & Rindskopf, 2007). In addition, it is informative to estimate the degree of variability in

the treatment effect estimate between participants (Shadish, Kyse & Rindskopf, 2013). It can be the case that a statistically significant average treatment effect is obtained across participants, but that there is a lot of variability in this estimate between participants, and that the treatment is not effective for all participants or even has adverse effects for some participants. In sum, inferences on the overall average treatment effect estimate need to be supplemented with an estimate of the between-participant variability because in single-case design studies the researchers care about individual participants (Barlow et al., 2009; Kazdin, 2011; Kratochwill et al, 2010, Kratochwill & Levin, 2014).

Van den Noortgate and Onghena (2003a, 2003b) proposed using a two-level regression model in order to capture in a single study both the overall average treatment effect across participants and between-participant variability in treatment effect estimates. This two-level regression model takes the hierarchical structure of the data into account, with observations nested within participants, and therefore takes into account that measurements from the same participant are more alike than measurements from different participants.

The two-level regression approach to summarize SCED data across participants using restricted maximum likelihood (REML) has been evaluated through a computer-intensive simulation study (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009), evaluating the relative parameter bias, relative standard error bias, the mean squared error, and the coverage proportion of the 95% confidence interval of the fixed effect estimates. Similarly, the bias and the precision of the variance components estimates have been evaluated. In their study, Ferron et al. (2009) found unbiased and precise fixed effect estimates. In contrast, the estimates of the variance components tended to be biased and imprecise (Ferron et al., 2009). Also subsequent research on the use of multilevel models to combine SCED data found that the variance component estimates

are unsatisfactory (e.g., Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b, 2013c). To avoid these problems, single-case researchers may want to turn to other estimation procedures such as Bayesian methods which seem to be promising for obtaining more precise variance components estimates (Baldwin & Fellingham, 2013). It makes sense to apply Bayesian estimation techniques in contexts of single-case experimental data because we deal with small sample sizes, so asymptotic assumptions are not likely to be met. Given the importance of obtaining unbiased and precise variance components estimates, this study was designed to evaluate the performance of Bayesian data analysis with a focus on variance components estimates.

Modeling of MBD Data

Two-level modeling allows for the estimation of the overall treatment effects over participants without losing information about participant-specific treatment effects (Van den Noortgate & Onghena, 2003a, 2003b; Shadish et al., 2013). Moreover we can estimate how much the treatment effects vary between participants within a study. Previous research focusing on estimation procedures for multilevel meta-analysis (e.g., Van den Noortgate & Onghena, 2003c) focused mainly on group-comparison designs and limited attention has been devoted to estimation procedures in the context of SCED studies (Kratochwill et al., 2010, Kratochwill & Levin, 2014).

To specify the two-level model that combines data across participants within an MBD, we can define regression equations at two levels. At the first level, we can use the following regression equation:

$$y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}D_{ij} + \beta_{3j}T_{ij}^{'}D_{ij} + \rho e_{i-1\,j} + e_{ij} \qquad (1)$$

y_{ij} indicates the continuous outcome score (e.g., math score), on measurement occasion *i* (*i* = 1,…*I*) for participant *j* (*j* = 1,…*J*) and is regressed on a dummy variable, D_{ij}, indicating the condition (e.g., 0 = baseline condition and 1 = treatment condition), a time variable, T_{ij} (i.e., to model a possible time trend during the baseline phase), and an interaction term between D_{ij} and the time variable T_{ij}' (i.e., because it is possible that the trend during the treatment condition will differ from the trend during the baseline condition). T_{ij}' is centered in a way that it equals zero at the start of the treatment phase, whereas T_{ij} equals zero at the start of the experiment. As a consequence, β_{0j} is the expected outcome score for participant *j* when T_{ij} and D_{ij} are zero, which means the expected score at the beginning of the baseline phase. β_{1j} is the time trend during the baseline phase; β_{2j} is the immediate treatment effect; and β_{3j} is the difference in trend between the baseline condition and treatment condition. For more details about coding the design matrix in contexts of SCEDs and the interpretation of regression coefficients, we refer the reader to Moeyaert, Ugille, Ferron, Beretvas, and Van den Noortgate (2014). The e_{ij}'s are residuals, which are usually assumed to be normally distributed around a mean of zero with a variance of σ_e^2, but other distributions are possible. If the outcome variable, y_{ij}, is non-continuous, for instance a count, the use of a Poisson or negative binomial distribution might be more appropriate (i.e., Breslow & Clayton, 1993; Shadish et al., 2013). ρ indicates the autocorrelation parameter. If ρ is positive, then errors closer in time are more similar; if ρ is negative, errors closer in time are more different and if ρ is zero, then there is no correlation between the errors. In SCED data, repeated measures are obtained within a person and as a consequence, the issue of autocorrelation cannot be neglected (Ferron et al., 2009; Huitema &

McKean, 1994; McKnight, McKean, & Huitema, 2000). Previous research indicates that not modeling existing autocorrelation in a two-level analysis of single-cases results in biased parameter estimates (Ferron et al., 2009). On the other hand, Shadish and Sullivan (2011) indicated that the size of autocorrelation in SSED studies varies tremendously (AR values across 809 studies ranged from -.931 to .786), with an average of .20. It is unlikely that the treatment effects are the same for all participants included in the SCED study. To capture the variability between participants and to estimate the average treatment effect across the participants, we allow the level-1 coefficients to vary at the second level.

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \\ \beta_{2j} = \theta_{20} + u_{2j} \\ \beta_{3j} = \theta_{30} + u_{3j} \end{cases} and \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & & & \\ \sigma_{u_0u_1} & \sigma^2_{u_1} & & \\ \sigma_{u_0u_2} & \sigma_{u_1u_2} & \sigma^2_{u_2} & \\ \sigma_{u_0u_3} & \sigma_{u_1u_3} & \sigma_{u_2u_3} & \sigma^2_{u_3} \end{bmatrix} \right) \quad (2)$$

The θ coefficients indicate the average effects over participants and the u's indicate how individual participants differ from this overall effect. These level-2 residuals are assumed to follow a multivariate normal distribution. Researchers, policy makers, and practitioners are mainly interested in the average immediate treatment effect over participants (θ_{20}) and the average treatment effect on the time trend over participants (θ_{30}), as well as in how much the effect vary over participants (c.q $\sigma^2_{u_2}$ and $\sigma^2_{u_3}$, indicating the between-participant variance of the immediate treatment effect and the treatment effect on time trend respectively). The between-participant variance can be found on the diagonal of the covariance matrix, whereas the off-diagonal elements refer to the covariance. Multilevel modeling estimation is complex because of the composite error structure, consisting of both level-1 and level-2 residuals (e_{ij}, u_{0j}, u_{1j}, u_{2j},

u_{3j}), and two kinds of parameters are to be estimated, namely the fixed effect coefficients and the residuals' (co)variance components. Assuming no autocorrelation (i.e., $\rho = 0$ in Equation 1), combining Equation 1 and Equation 2 results in Equation 3:

$$y_{ij} = \theta_{00} + u_{0j} + (\theta_{10} + u_{1j})T_{ij} + (\theta_{20} + u_{2j})D_{ij} + (\theta_{30} + u_{3j})T'_{ij}D_{ij} + e_{ij} \quad (3)$$

Level-1 residuals are assumed to be independent of level-2 residuals: Cov (e_{ij}, u_{0j}) = Cov (e_{ij}, u_{1j}) = Cov (e_{ij}, u_{2j}) = Cov (e_{ij}, u_{3j}) = 0.

Estimation Methods in the Two-level Modeling of SCED Data

Maximum Likelihood Estimation

Maximum likelihood (ML) algorithms are the default in most statistical software programs for multilevel analysis. For an introduction to the ML estimation algorithm, we refer the reader to Goldstein (1995), Raudenbush and Bryk (2002), and Snijders and Bosker (2012). ML estimates have desirable large sample properties: (1) they are consistent (i.e., as the sample size increases, the ML estimates tend to approach the true parameter value), (2) they are asymptotically normal (i.e., the ML estimates will have an approximate normal distribution centered around the true parameter value), which simplifies significance testing and the construction of confidence intervals for the parameters, and (3) the estimated likelihood function can be used for assessment of the model fit and comparison among models. ML parameter estimates are the parameter values that make the data most likely (that maximize the likelihood). The difference between full ML and restricted ML (FML and REML) lies in the way the (co)variances are estimated. When using REML, the principle of ML is applied to the least-squares residuals. This means that the (co)variances are estimated after controlling the observed scores for the fixed effects (Harville, 1977; Patterson & Thompson, 1971; Robinson, 1991; Thompson, 1980). An advantage is that the (co)variance component estimates are less biased.

However, a drawback of REML is that the deviance scores (i.e., minus 2 times the residual log likelihood) cannot be used for comparing models that also differ in their fixed part, because the data then are corrected for two different fixed parts. The difference between both approaches becomes especially visible when a small number of level-2 units are included.

In the context of SCED studies, we deal with small sample sizes and so asymptotic assumptions are rarely met. The central limit theorem cannot be applied and consequently the validity of the statistical inferences cannot be based on this theorem. Previous simulation studies studying the use of two-level models to summarize SCED data indicate that the FML and REML estimates for the fixed effects and the corresponding standard errors are still unbiased (Ferron et al., 2009). This stands in contrast to the estimates of the variance components, which tend to be biased and imprecisely estimated. This bias in variance estimates may result in biased standard errors and therefore flawed inferences on fixed effects (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b). In the past, adjustments have been suggested to ML methods for inferences on the fixed effects estimates (Kenward & Roger, 1997, 2009). However, these adjustments do not deal with the uncertainty about the variance components estimates (Baldwin, Murray et al., 2011, Burrick & Graybill, 1992, Kenward & Roger, 1997, 2009). For this reason, we discuss the Bayesian estimation approach as an alternative to ML estimation, as it incorporates uncertainty about variance components in the estimation of fixed effects.

Bayesian Estimation

It is natural and logical to apply Bayesian estimation[1] in the context of SCEDs because is it not based on asymptotic assumptions (de Vries & Morey, 2013; Rindskopf, 2014).

[1] Fully Bayesian estimation is different from Empirical Bayes estimation. Empirical Bayes estimation can be used to get participant-specific treatment effect estimates, using the empirical data of all participants as a prior. The estimates are also called shrinkage estimates, because the estimates are in general closer to the mean estimates

Conceptually, Bayesian inference is simple: prior beliefs about one or more parameters are expressed in a statistical model, and observed evidence is used to update these prior beliefs. Bayes' theorem is used to combine prior beliefs with observed evidence (i.e., data), producing the probability of a parameter given the data (Spiegelhalter, Abrams, & Myles, 2004):

$$P(\theta|y) = \frac{P(y|\theta)\,p(\theta)}{p(y)} \quad (4)$$

where $P(y|\theta)$ is the probability of the observed data for each possible value of θ (i.e., the likelihood), $p(\theta)$ is the prior distribution of θ, that is, a distribution that represents knowledge about the parameter prior to observing the data (based on previous research in the field and/or experts' knowledge), and $p(y)$ is the probability of the data. The left hand side of Equation 4, $P(\theta|y)$, indicates the posterior distribution and Bayesian estimation consists of finding this posterior distribution of unknown parameters based on the observed data and a prior distribution defined by the researcher. The construction of the posterior distribution can be complex and is not always analytically feasible, such as in the context of multilevel modeling, because integration over high-dimensional probability distributions is needed. Markov Chain Monte Carlo (MCMC) simulation methods can be very helpful in this context. MCMC is a general method based on drawing values of a parameter of interest from the target distribution. MCMC is an iterative process. Once the joint target posterior distribution is reached, we expect to draw from the posterior distribution at each step of the process. There are many ways to construct the Markov Chains, but all of them are special cases of the general framework of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and Hastings (1970). After a specific number of iterations, the chain has reached its target distribution and the samples that are drawn up to

---

across participants than when using the individual participant data only. In contrast, in full Bayesian estimation, priors are defined in advance, and incorporated in the analysis.

that point (in what is called the 'burn-in period') can be thrown away. The minimum period of the burn-in is dependent on the model complexity and on certain characteristics of the model (e.g., correlation among parameters). An important aspect of Bayesian analysis is checking for convergence of the Markov Chains. Common approaches to assessing convergence include examining trace plots and diagnostic statistics, such as the Geweke diagnostic, the Heidelberger-Welch test of stationarity, and the Gelman-Rubin diagnostic (Jackman, 2009; Gelman, Carlin, Stern, and Rubin , 2013).

The posterior distribution is used to make probability statements. For instance, if the parameter of interest is θ_{20} (i.e., the overall average immediate treatment effect), the posterior distribution allows one to make direct statements such as "the probability that θ_{20} exceeds 0 is 95%". The posterior distribution also allows for calculating credible intervals. These intervals are somewhat comparable to the usual confidence intervals from a frequentist (e.g., FML or REML) approach, although they are conceptually different. A 95% confidence interval from the frequentist approach is interpreted as follows: if we repeated this study many times, 95% of the confidence intervals computed using this method would contain θ_{20}. In contrast, a credible interval is interpreted directly in terms of probability: there is a 95% probability that the parameter value is in the 95 % credible interval. Despite the conceptual differences, we will compare in this study 95% credible intervals from the Bayesian approach and 95% confidence intervals from the frequentist approaches.

It is important to notice that in general Bayesian inference is conceptually different from ML inferences. In traditional frequentist theory, a null hypothesis is assumed and the likelihood of observing the data y given the model, [i.e., $P(y|\theta)$], is tested. If given the assumed model it is unlikely to observe data that are at least as extreme as the data at hand (i.e., $p < .05$), then we

reject the null hypothesis (note that this gives no information about the likelihood that the assumed model is correct). In the Bayesian approach we see the reverse: the assumed model is rejected as being unlikely given the data. For a more in depth introduction to Bayes' theorem and Bayesian computation, we refer the reader to Carlin and Louis (2009), Gelman et al. (2013), and Lynch (2007).

Selecting Priors

Previous methodological research devoted to Bayesian statistics within social, educational, and behavioral sciences is rather scarce and is situated within growth mixture modeling (Depaoli, 2014; 2015), growth curve models (Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2007), multilevel modeling involving large sample sizes (e.g., Browne, Draper, Goldstein, & Rasbash, 2002) and partially clustered data with small sample sizes (e.g., Baldwin & Fellingham, 2013). Limited methodological research within the field of two-level modeling of SCEDs has been conducted, even though experts within the field have advocated for this alternative estimation approach (e.g., Rindskopf, 2014; Shadish et al., 2013).

As can be deduced from Equation 3, a total of nine parameters need to be estimated, if for simplicity we assume that the u's are independent (i.e., covariance is zero): four fixed effects (θ_{00} , θ_{10} , θ_{20} , and θ_{30}) and five random effects ($\sigma^2_{u_0}$, $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, $\sigma^2_{u_3}$ and σ^2_e).

Prior Selection for the Fixed Effects and the Within-Participant Variance. Previous research is consistent in recommending default non-informative priors for the fixed effects and the within-participant variance in multilevel modeling contexts because this choice results in unbiased and precise estimates (Gelman, 2006, Gelman et al., 2013, Spiegelhalter et al., 1994, 2003). Based upon Gelman (2006) and Gelman et al. (2013) a non-informative normal

distribution with a mean of 0 and a variance of 10^6 can be chosen as prior for each of θ_{00} , θ_{10} , θ_{20} , and θ_{30} :

$$\theta_{00}\ ,\ \theta_{10}\ ,\ \theta_{10}\ ,\ \theta_{30}\ \sim \text{Normal}(0,\ 10^6) \tag{5}$$

Based upon Spiegelhalter et al. (1994, 2003), an inverse gamma prior (with shape and scale parameters each set to a value of 0.001 and 0.001 respectively) can be used for the level-1 residual variance parameter:

$$\sigma_e^2 \sim inverse-gamma(0.001, 0.001) \tag{6}$$

Prior Selection for the Between-Case Variance Components. The literature is less consistent about which priors to define for the level-2 variance components (Gelman, 2006). The choice of a prior distribution for level-2 variance components can have a substantial impact on inferences, especially in the case when the number of level-2 units is small or the corresponding level-2 variance is small (Gelman, 2006; Gelman et al., 2013). Therefore, care needs to be given to the prior selection for the between-case variance components. Constructing default non-informative priors is not a viable option as previous research indicates that this results in biased and unprecise estimates. Gelman (2006) used a uniform distribution with lower limit of 0 and a large value for the upper limit such as 50 or 100 as a non-informative prior. Gelman (2006) found that this prior might lead to positively biased and less accurate estimates if there are a very limited number of units (equal to or less than 3) and the population variance is expected to be small.

In sum, it seems reasonable to use weakly informative priors. We use the term *weakly informative priors* (Gelman, 2006) when the information they contain is intentionally weaker than the actual prior knowledge that is available. The use of weakly informative priors has several advantages: (1) This category of priors have the potential to contain some information to

regularize the posterior distribution; that is to keep it within reasonable bounds but without attempting to fully capture ones scientific knowledge about the underlying parameter; (2) Gelman et al. (2013) state that the choice of non-informative prior distributions can have a big effect on inferences, especially for problems where the number of groups is small (equal to or less than 3) or the group variance is small (which are realistic conditions for SCEDs); (3) By choosing priors that are only weakly informative, we can still let to speak the data for themselves (Spiegelhalter et al., 2004).

Spiegelhalter et al. (2004) advise to analyze past hierarchical models in similar contexts to determine reasonable values for the between-participant variance. As a consequence, re-analyses of published meta-analyses of SCEDs in a given research area can be conducted to identify reasonable expected values for the variance components (Lau, Schmid, & Chalmers, 1995; DerSimonian, 1996). In order to suggest reasonable priors, we conducted such re-analyses. We retrieved raw SCED data from graphs displayed in primary SCED studies (for more details about the SCED data retrieval process, we refer the reader to Moeyaert, Maggin, and Verkuilen, 2016[2]) included in five meta-analyses, all investigating the effectiveness of treatments to reduce problem behavior for students with special needs (Denis, Van den Noortgate, and Maes, 2011, Kokina and Kern, 2010; Shogren, Fagella-Luby, Bae, and Wehmeyer, 2004; and Wang, Cui, & Parrila, 2001). To make data from these meta-analyses comparable, we standardized the raw data (as described by Van den Noortgate & Onghena, 2008). The results of these re-analyses can be found in Table 1.

--- INSERT TABLE 1 ABOUT HERE ---

[2] An alternative is to contact the authors of the published meta-analysis to request the raw data. However, consistent with other fields of meta-analysis, researchers have found that the response rate associated with requests for the original data values from authors of original studies tend to be prohibitively low (Shadish & Sullivan, 2011; Manolov & Solanas, 2013).

Based on these results, and based on previous recommendations, three different types of weakly-informative prior distributions (with different degrees of information) are considered for the level-2 residuals' standard deviation or variance. The first class of prior distributions are the half-Normal distributions, which are relatively easy to interpret as we are familiar with the normal distribution. A second set of prior distributions that will be investigated are half-Cauchy distributions, suggested by Gelman (2006) in scenarios for conditions representing a small number of level-2 units. Another typical distribution chosen for variances and covariance matrices are the inverse-Wishart distributions (e.g., Gelman & Hill, 2007), which are the conjugate prior for the covariance matrix of multivariate normal distributed variables, which implies that when it is combined with the likelihood function, it will result in a posterior distribution that belongs to the same distributional family. These three classes of prior distributions are discussed in detail in the next sections and an overview of the prior distributions can be found in Table 2.

--- INSERT TABLE 2 ABOUT HERE ---

Half-Normal Prior Distributions The first class of plausible and simple prior distributions for the standard deviations across participants are half-Normal distributions characterized by positive values only and values closer to zero having a larger likelihood to occur than values further away from zero. The value for the variance is defined based upon prior knowledge (see Table 1) and defines the spread of the distribution. We choose for the half-Normal instead of the normal distribution as the variance (and standard deviation) is constrained to be positive: Half-Normal $\sim$ [0, $(SD_u/1.96)^2$] (Pauler & Wakefield, 2000). This distribution has its mode at 0 and is steadily declining in standard deviation, with Percentile 95 being the SD_u. Its median will be $\phi^{-1}(0.75) \times SD_u/1.96 = 0.39\ SD_u$. We will explain the logic for

including prior information in the distributions for the between-case standard deviation of the immediate treatment effect (σ_{u_2}), but the same reasoning can be applied for the between-participant standard deviation for other parameters' residuals ($\sigma_{u_0}, \sigma_{u_1}, \sigma_{u_3}$). Based upon the re-analyses of meta-analyses, we found an upper limit of 4.980 for the between-case variance of the immediate treatment effect, corresponding to a standard deviation of 2.23. Because we wanted to make the prior less informative, we choose the following upper limits for the SD (SD_u): 6, 9 and 14. The following formula, using SD_u, can be applied to specify the variance for the half-Normal: $(SD_u/1.96)^2$, resulting in the following values: 10 [~$(6/1.96)^2$], 20 [~$(9/1.96)^2$] and 50 [~$(14/1.96)^2$]. The medians of the distributions will be 2.34 (0.39 × 6), 3.51 (0.39 × 9), and 5.46 (0.39 × 14) respectively. A graphical display of the half-Normal distributions is given in Figure 2.

--- INSERT FIGURE 2 ABOUT HERE ---

Half-Cauchy prior distributions. The second class of prior distributions for standard deviations are the half-Cauchy distributions characterized by positive values and values closer to zero having a larger likelihood to occur than values further away from zero. In contrast to the half-Normal distribution, the shape of the distribution is defined by a location parameter (i.e., where the peak of the distribution is located) and a scale parameter (which specifies the half-width at half-maximum). We chose the half-Cauchy instead of the Cauchy distribution as the variance (and standard deviation) is constrained to be positive. The half-Cauchy can be a convenient weakly informative family; the distribution has a broad peak at zero and a single scale parameter. In the case that B (scale parameter) is very large (B → ∞) this becomes a uniform prior density. Large but finite values of B represent prior distributions which we consider weakly informative because, even in the tail, they have a slope (unlike for example a

half-Normal distribution) and can let the data dominate if the likelihood is strong in that region. In a similar situation (3 participants and small variance), Gelman (2006) used a half-Cauchy prior distribution on the standard deviation with a scale parameter B = 25 (a value chosen to be a bit higher than expected for the standard deviation of the underling fixed effects), so that the model will constrain the standard deviation only weakly. As a consequence, the prior distribution is high over the plausible range falling off gradually beyond that point removing much of the unrealistic upper tail. Following this logic, based upon the re-analyses of meta-analyses, the maximum between-case variance for the immediate effect was found to be 4.980 (the standard deviation 2.23). Therefore, we choose scale parameter larger than this values, being B = 10, 20 or 50 (these values are more or less 5 times, 10 times and 20 times larger than we expect for the between-case standard deviation of the treatment effect. As is clear from the figures (see Figure 3): the larger the scale parameter, the less informative the prior.

--- INSERT FIGURE 3 ABOUT HERE ---

Inverse-Wishart distributions. The third and last class of prior distributions under investigation in this study are the inverse-Wishart distributions. This class of distributions can also be applied in multivariate multilevel models. The inverse-Wishart distribution is a multivariate generalization of the scaled inverse χ^2 to describe the prior distribution of the variance covariance matrix. The Wishart prior with small degrees of freedom and a fixed scale matrix is commonly used as a reference (non-informative) proper prior. Setting $\Sigma \sim iwish_{d+1}(I)$ has the appealing feature that each of the correlations in Σ have, marginally, a uniform prior distribution. Σ refers to the covariance matrix, d refers to the dimensions of the covariance matrix (four in current study) and I refers to the identity matrix (diagonal elements of the scale matrix are set to 1 and the off-diagonals are set to zero). The degrees of freedom needs to be

larger than the dimension of the matrix. As a consequence the degrees of freedom needed to be at least five and so we choose to set the degrees of freedom to six. In a pilot study, we varied the degrees of freedom to investigate the influence on parameter recovery, but no differences were observed. In addition to specifying the scale matrix using the identity matrix, we suggest a second prior specification: one in which the scale matrix is based on prior estimates of the variances of the random parameters. For instance, the mean/median values of the between-case variances of the re-analyses of meta-analyses can be used. An additional advantage of the inverse-Wishart distribution is that it ensures positive definiteness of the covariance matrix. Because the inverse-Wishart distributions are multidimensional distributions describing variance/covariance matrices, they are more difficult to visualize (Tokuda, Ben Goodrich, Van Mechelen, Gelman & Tuerlinckx, 2011). Therefore, Figure 4 displays the scaled inverse χ^2 distribution with six degrees of freedom, and varying the scale parameter. The scale parameter is the inverse of the variance. When we use the identity matrix, the scale parameter is 1. When we use the non-identity matrix and the variance is set to 2, the scale matrix is 1/2. As is clear from Figure 4, the smaller the scale parameter (larger the variance), the less informative the prior.

--- INSERT FIGURE 4 ABOUT HERE ---

Simulation Study

The purpose of current study is to compare the performance of ML estimation and Bayesian estimation in contexts of small sample sizes as no previous research is conducted in this context. In addition, we evaluate parameter recovery in the Bayesian approach for three classes of plausible weakly informative prior distributions (half-Normal, half-Cauchy and inverse-Wishart, with characteristics as described above). The influence of the functional form and the level of information included in the prior specification on the parameter recovery has not

been studied up to date. We expect the half-Cauchy prior distributions to outperform the half-Normal, because it allows for occasional larger standard deviations while still performing a reasonable amount of shrinkage towards zero. In other words, we think the set of true standard deviations that we might encounter has a distribution less like a normal than like a Cauchy with many small values and occasional large ones (Spiegelhalter et al., 2004). We expect the inverse-Wishart resulting in less good results compared to the half-Normal and the half-Cauchy as it has problems similar to the inverse gamma for variances (Gelman, 2006). The inverse gamma prior distribution (ϵ, ε) is an attempt at noninformativeness within the conditionally conjugate family, with ϵ set to a low value: 1, 0.01 or 0.001. A difficulty of this prior distribution is that in the limit $\epsilon \to 0$ it yields an improper posterior density, and thus ϵ must be set to a reasonable value. As recognized by Daniels and Kass (2001), in small samples the specification of the scale matrix can be influential.

In addition to this basic simulation study, we investigate two extensions for a subset of design conditions, namely (1) the modeling of autocorrelation and (2) the inclusion of a more informative prior distribution.

Simulating Two-Level Data

To compare the performance of Bayesian estimation and ML estimation to analyze two-level SCED data, we simulated raw data using Equation 3 in SAS (version 9.4, SAS Institute Inc., 2011-2014). Previous research recommends using REML over FML in contexts of SCEDs, because of the small sample sizes (Ferron et al., 2009; Owens & Ferron, 2012). However, we will include both FML and REML for completeness in the current simulation study, because no previous research in the context of SCEDs compared the variance estimates under both estimation procedures. In order to simulate realistic SCED data, we conducted several re-

analyses of SCED meta-analyses (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011; Table 1). The conditions chosen for the simulation study were also based on a review of SCED studies conducted by Shadish and Sullivan (2011), Farmer et al. (2010) and Ugille et al. (2015). We varied four design conditions, namely the number of measurements within a participant (*I)*, the number of participants within an MBD (*J*), the immediate treatment effect (θ_{20}), and the treatment effect on the time trend (θ_{30}). We chose the following parameter values for the design conditions: number of level-1 units (measurement occasions)*: I* = 20 or 40, and number of level-2 units (participants): *J* = 3, 5 or 7. These numbers were selected based on the survey of SCED studies of Shadish and Sullivan (2011), where the number of cases per study ranged from 1 to 13 with median 3, and on a review of Farmer, Owens, Ferron and Allsopp (2010), where 93% of the average number of cases per study fell at or below 7. Because we focus on the MBD across participants design (an example is displayed in Figure 1), we simulated the data in a way that the start of the treatment is staggered across participants as depicted in Table 3.

--- INSERT TABLE 3 ABOUT HERE ---

For instance, if three participants are simulated with each participant having 20 measurements, the start of the treatment for the participants occurs at the seventh, eleventh, and fifteenth time points. The immediate effect of the treatment (θ_{20}), was varied to have values of 0 or 2 and the treatment effect on the time trend (θ_{30}) was set to 0 or 0.2. Because the interest lies in the estimate of the treatment effects (θ_{20} and θ_{30}), the initial baseline level (θ_{00}) and the trend during the baseline (θ_{10}) were kept to a constant value of 0 for simplicity. Covariances between regression coefficients were set to zero and therefore the between-participant covariance matrix

is a diagonal matrix: $diag(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u3}) = diag(2, 0.2, 2, 0.2)$. σ^2_{u0}, and σ^2_{u1}, $\sigma^2_{u_2}$ and $\sigma^2_{u_3}$ indicate the between- participant variance of the initial baseline level, the trend during the baseline, the immediate treatment effect and the treatment effect on the time trend respectively. The reason for setting the covariances to zero is to avoid too much complexities in the current simulation study. For the basic simulation study, the level-1 variance was kept to a constant value of 1.

Analyzing the Simulated Two-Level Data

A total of 8,000 datasets were simulated using Equation 3, 500 datasets for each of the 16 conditions. The datasets were analyzed ten times: using FML, using REML and using the Bayesian framework with eight different weakly informative priors. These eight prior distributions were discussed earlier in the section "choosing priors" (see Table 2 for an overview).

For the analyses using ML estimation, PROC MIXED in SAS (version 9.4, SAS Institute Inc., 2011-2014) was used. The Kenward-Roger method (Kenward & Roger, 2009) was used to estimate the degrees of freedom because previous research has indicated that this method to adjust the standard errors and degrees of freedom resulted in unbiased fixed effect estimates, and in accurate confidence intervals for the fixed effects (Ferron et al., 2009). For Bayesian estimation, PROC MCMC from SAS version 9.4 was used. A preliminary simulation study indicated that for all included conditions the Markov Chain was stable after a burn in period between 5,000 and 10,000 iterations. This was evaluated by looking at the diagnostic tests (Geweke diagnostic, Heidelberger-Welch test of stationarity, and the Gelman Rubin diagnostic) and visual inspection of the trace plots for one simulation set for the 16 conditions. In order to ensure convergence in the final simulation, we set the burn in to 10,000. After the burn-in period

and thinning the distribution by keeping every 25th simulated draw from each sequence, 100,000 samples were taken from this distribution and the central tendency of the posterior distribution is given (i.e., including the posterior median, referring to the parameter's point estimate). Thinning is a common strategy for reducing sampling autocorrelations. For the fixed effect estimates, the prior probability distribution and the posterior distribution are in the same distributional family (they are called conjugate distributions) and as a consequence the Gibbs sampler is used to draw samples. This is also the case for the inverse-Wishart distribution for the variance components. Conjugate sampling is efficient, because it enables the Markov Chain to obtain values from the target distribution directly. The half-normal and half-Cauchy prior and posterior distribution for the standard deviations are not in the same family and as a consequence, the Metropolis-Hastings sampling method is used.

In order to evaluate the performance of the ten analysis procedures (FML, REML and the Bayesian approach using eight different priors), we calculated the (relative) bias, the mean squared error, and the coverage proportion of the 95% confidence/credible interval of the estimated model parameters of interest, namely the immediate treatment effect ($\hat{\theta}_{20}$), the treatment effect on the time trend ($\hat{\theta}_{30}$), and the between-participant standard deviation of the treatment effects (i.e, $\hat{\sigma}_{u_2}$ and $\hat{\sigma}_{u_3}$).

The (estimated) absolute bias is the difference between the average effect estimate (the mean of the estimates across all replicated datasets) and the true population value. The relative bias is the absolute bias divided by the true population value. Therefore, the relative bias can only be estimated for non-zero true population values. For instance, the relative bias of the immediate treatment effect is: $(\bar{\hat{\theta}}_{20} - \theta_{20})/\theta_{20}$ .For the variance components estimates, between-

case standard deviation estimates are obtained when the half-Cauchy and the half-Normal distributions are used as priors. Therefore, the relative bias of the between-case standard deviation is obtained by: $\bar{\hat{\sigma}}_u - \sqrt{\sigma_u^2} / \sqrt{\sigma_u^2}$. Inverse-Wishart priors are specified on the variances and as a consequence between-case variance estimates are obtained. In order to make the scale of the relative bias comparable, we first calculated the between-case standard deviation ($\sqrt{\bar{\hat{\sigma}}_u^2}$) and used this to calculate the relative bias of the estimated between-case standard deviation: $\sqrt{\bar{\hat{\sigma}}_u^2} - \sqrt{\sigma_u^2} / \sqrt{\sigma_u^2}$. We conclude that a parameter estimate is unbiased if the relative bias is smaller than the 5% cut off criterion set by Hoogland and Boomsma (1998). As preliminary analysis, we compared the absolute bias across all conditions. As no clear differences were found between the $\theta_{20} = 0$ and $\theta_{20} = 2$ conditions and between the $\theta_{30} = 0$ and $\theta_{30} = 0.2$ conditions , we choose to report only the results for the $\theta_{20} = 2$ and $\theta_{30} = 0.2$ conditions allowing us to present the relative bias instead of the absolute bias, and therefore to use the cut off criterion set by Hoogland and Boomsma (1998).The mean squared error (*MSE*) is defined as the mean squared difference between the estimates and the population value, which gives important information about both the bias and the variance of the estimates (*MSE* = bias^2 + variance). We want the relative bias and the *MSE* to be as low as possible. In addition, we evaluate the coverage proportion of the 95% confidence or credible interval (*CP95*). This can be accomplished by constructing 95% confidence/credible intervals around the effect estimates (i.e., fixed effects and random components) and calculate in what proportion of the replicated datasets the 95% confidence/credible interval contains the true population value. We hope the *CP95* is close to the nominal level of .95. Because we simulated 500 datasets for each condition, the coverage proportions of 95 % confidence intervals can be estimated relatively accurately. More

specifically the standard error can be calculated using the following formula: $SE(p) = \sqrt{[p(1-p)]/B}$, in which p stands for the probability, being .95 and B the number of replications, being 500. As a consequence, the expected standard error for a probability of .95 is 0.010 $\left(= \sqrt{(0.95 * 0.05)/500}\right)$ and therefore, we expect that 95 % of the coverage proportions of the 95 % confidence intervals range from 93.04 % to 96.96 % (95 % ± 1.96 *1 %).

Constructing confidence intervals around the variance components is less straightforward because their distribution is extremely skewed and bounded at zero. Therefore, SAS Proc MIXED uses a Satterthwaite approximation that has a lower boundary constraint of zero. The formula that SAS 9.3 uses to calculate the lower and upper limits is displayed in Equation 7.

$$\frac{df' \times \hat{\sigma}_{\mathrm{u}}^2}{\chi^2_{df',.975}} \leq \sigma_u^2 \leq \frac{df' \times \hat{\sigma}_{\mathrm{u}}^2}{\chi^2_{df',.025}} \quad (7),$$

with df' indicating the adjusted degrees of freedom, $\hat{\sigma}_{\mathrm{u}}^2$ the estimated level-2 variance, $\chi^2_{df',.975}$ the lower critical chi-square with df' degrees of freedom and $\chi^2_{df',025}$ the upper critical chi-square value with df' degrees of freedom. The adjusted degrees of freedom, df' is calculated as 2 times the square of the Wald Statistics (i.e., $2 \times [\hat{\sigma}_u^2 / SE(\hat{\sigma}_u^2)]^2$. However, in the context of this study, this can result in unrealistically large upper limits.

The formula for the upper limit is very sensitive to the degrees of freedom. For one degree of freedom, the divisor is 0.001, for two degrees of freedom it is 0.056, for three degrees of freedom it is 0.216 and for four it is 0.484. The divisor changes by a factor of almost 500 for a change from one to four degrees of freedom. The lower limit only doubles for a change from one to four degrees of freedom. Therefore we propose an adjustment to calculate the upper limit of the 95% confidence interval for variance estimates. Instead of using df' in the right part of Equation 4 we used the degrees of freedom equaling to the total number of participants – 1 and

then apply Equation 4. Using the total number of participants – 1 is the same as using the degrees of freedom for an ordinary variance, which might not be right, but it is closer than what is expected than the Satterthwaite approximation. The Satterthwaite approximation is only accurate in certain circumstances, and has not been tested for small samples (SAS Institute Inc., 2011-2014).

In order to study the variation in relative bias, *MSE*, *CP95* of the estimated treatment effects and variance components of interest, we used PROC GLM (i.e., analysis of variance, ANOVA) in SAS 9.4. We investigated whether the estimates are dependent on the analysis procedure[3], design factors, and (two-way) interactions of these. Because assumptions underlying the ANOVAs, such as normality and homoscedasticity, are questionable, the ANOVA procedure was only used as preliminary analyses, with the results only interpreted as giving a primary indication. We did not only look at the statistically significance of main and interaction effects ($p < .0001$), but we also calculated the eta-squares (i.e., $\hat{\eta}^2$) as effect sizes indicating whether the estimated main effects and/or interaction effects are rather small (.02), medium (.13) or large (.26; Cohen, 1988).

The SAS code for multilevel modeling and Bayesian estimation is included as supplementary material.

Results of the Simulation Study

Because of space limitations, we only display the results for $\theta_{20} = 2$ and $\theta_{30} = 0.2$. The same patterns are obtained when the population treatment effects are set to 0. The full results can be requested from the first author.

Results for Average Estimated Treatment Effects

[3] The estimation procedure (i.e., ML procedures and Bayesian procedures) are included as a within-subjects factor in the analysis of variance.

Relative bias and mean squared error. The results of the fixed effect estimates (i.e., $\hat{\theta}_{20}$ and $\hat{\theta}_{30}$) are in line with the results of previous research about multilevel modeling that documents unbiased (i.e., relative bias < .05) and precise estimates using the FML, REML (Moeyaert et al., 2013b, 2013c), and Bayesian estimation using a non-informative prior normal distribution with a mean of 0 and a variance of 10^6 on the fixed effects (Gelman, 2006, Gelman et al., 2014). In terms of the bias, no large difference between the ten analyses procedures were found for both bias of $\hat{\theta}_{20}$,$F(9, 27577) = 0.44$, $p = .920$, $\hat{\eta}^2 = .0001$ and bias of $\hat{\theta}_{30}$, $F(9, 27557) = 0.02$, $p = 1.000$, $\hat{\eta}^2 < .0001$.

Also for the *MSE*, no significant large differences were found between the procedures: for the *MSE* of $\hat{\theta}_{20}$, $F(9, 27577) = 0.54$, $p = .850$, $\hat{\eta}^2 = .0002$, and for the *MSE* of $\hat{\theta}_{30}$, $F(9, 27557) = 0.80$, $p = .621$, $\hat{\eta}^2 = .0002$. The number of participants contributed significantly to the precision of the estimates: for $\hat{\theta}_{20}$, $F(2, 27557) = 936.68$, $p < .0001$, $\hat{\eta}^2 = .063$ and for $\hat{\theta}_{30}$, $F(2, 27557) = 855.40$, $p < .0001$, $\hat{\eta}^2 = .058$. The same is true for the number of measurements within participants: $F(1, 27557) = 40.71$, $p < .0001$, $\hat{\eta}^2 = .0013$ for $\hat{\theta}_{20}$ and $F(1, 27557) = 98.42$, $p < .0001$, $\hat{\eta}^2 = .0033$ for $\hat{\theta}_{30}$. The larger the number of measurements and especially the number of participants, the more precise the estimates for $\hat{\theta}_{20}$ and $\hat{\theta}_{30}$. This is in line with previous research findings in multilevel contexts suggesting that the number of units at the highest level are most important in order to get unbiased and precise estimates (Hox, 2002).

Coverage proportion of the 95% confidence interval. For the coverage proportion of the 95% confidence interval (i.e., *CP95*), we found a statistically significant and large main effect of the analysis procedure [$F(9, 18) = 174.12$, $p < .0001$, $\hat{\eta}^2 = .818$] and a statistically

significant interaction effect between the number of second level units and the analysis procedure for $\hat{\theta}_{20}$ [$F(18,18) = 16.38$, $p < .0001$, $\hat{\eta}^2 = .002$].

A first finding is that the ML procedures resulted in a too small *CP95* when only three participants were included. By increasing the number of participants from 3 to 5, this problem was solved.

A second finding is that the Inverse-Wishart distribution using the identity matrix resulted in a *CP95* that was too low across all conditions (see Model 9 in Figure 5). In contrast, the inverse-Wishart using the non-identity values resulted in a *CP95* close to the nominal level, but was slightly underestimated when the number of observations was set to a larger value (I = 40) in combination with five or seven participants.

A third finding is that the other Bayesian priors result in an appropriate *CP95* when the number of participants was set to 5, with having at least 40 measurement occasions. If the number of participants was set to 7, the *CP95* values were acceptable, independent of the number of measurements.

These three findings are graphically represented in Figure 5. Similar patterns were found for $\hat{\theta}_{30}$.

--- INSERT FIGURE 5 ABOUT HERE ---

Results for Variance Components

Relative bias and mean squared error. The ANOVA indicated a statistically significant main effect of the analysis procedure [$F(9, 27558) = 204.06$, $p < .0001$, $\hat{\eta}^2 = .061$] and the number of participants [$F(2, 27558) = 50.53$, $p < .0001$, $\hat{\eta}^2 = .003$] on the relative bias of the between-participant standard deviation of the immediate treatment effect. However, the $\hat{\eta}^2$

values indicated that this effect was rather small. When including more participants, the relative bias became smaller as well as the difference between ML and Bayesian procedures (see Figure 6).

--- INSERT FIGURE 6 ABOUT HERE ---

When looking in more detail, a first finding is that the ML procedures underestimated the between-participant standard deviation of the immediate treatment effect. The relative bias was consistently larger in magnitude than 5% and became smaller as the number of participants increased. For instance, for the FML, the relative bias ranged from -.54 (when there were only three participants and 20 measurement occasions) to -.15 (when there were seven participants having 40 measurements). The REML resulted in less biased estimates compared to the FML and ranged from -.33 (when there were only three participants and 20 measurement occasions) to -.06 (when there were seven participants and 40 measurement occasions).

A second conclusion is that the inverse-Wishart prior distributions consistently resulted in negatively biased estimates (i.e., relative bias was larger than – 5 % across all conditions). The relative bias for the inverse-Wishart with the non-identity values was smaller compared to the inverse-Wishart with the identity matrix, but is still larger than the threshold level of 5% (with a range from -.11 to -.16).

As for the other prior distributions, we found positive values for the relative bias. The degree of relative bias was dependent on the analysis procedure when only three participants were included (independent of the number of measurements). In these conditions, the relative bias using half-Normal distributions was smaller compared to half-Cauchy distributions. The half-Normal distribution containing most information [i.e., half-Normal ~ (0, 10)] resulted in the smallest bias (i.e., relative bias = .19) and the relative bias increased when less prior information

was included. The smallest relative bias for the Half-Cauchy in this condition was .39. When increasing the number of participants from three to five, the relative bias for all models, except from the Wishart models, decreased and only slightly biased estimates were obtained (see Figure 6). By further increasing the number of participants (from five to seven) no additional decrease in relative bias was found. Similar results were obtained for the estimate of the between-participant standard deviation of the treatment effect on the trend ($\hat{\sigma}_{u_3}$).

We also investigated how precise the between-participant standard deviations were estimated using the ten different analysis procedures by calculating the *MSE*. The preliminary ANOVA indicated a statistically significant main effect of the analysis procedure [$F(9, 27558) = 204.06$, $p < .0001$, $\hat{\eta}^2 = .061$] and the number of participants [$F(9, 27558) = 50.53$, $p < .0001$, $\hat{\eta}^2 = .0034$].

--- INSERT FIGURE 7 ABOUT HERE ---

Increasing the number of participants resulted in a more precise between-participant standard deviation estimate; this applies for all analysis procedures except from the inverse-Wishart distribution using the identity matrix. When the number of participants was small (i.e., $J = 3$.), large differences in *MSE* of $\hat{\sigma}_{u_2}$ between the analysis procedures were found (independent of the number of measurements). In these conditions, the *MSE* was smallest for the ML procedures. However, when the number of participants were set to 5 or 7 no differences between the ML and Bayesian procedures (except for the Inverse-Wishart distribution with the identity matrix) were observed. These results are graphically displayed in Figure 7. Similar results were obtained for $\hat{\sigma}_{u_3}$.

Coverage proportion of the 95% confidence interval. We found a statistically significant large main effect of the analysis procedure [$F(9, 18) = 104.27$, $p < .0001$, $\hat{\eta}^2 = .700$]

and a statistically significant small effect of the number of participants [$F(9, 18) = 11.60$, $p = .0006$, $\hat{\eta}^2 = .017$] on the *CP95* of the between-participant standard deviation of the immediate treatment effect. An interaction between the analysis procedure and the number of participants was found [$F(18, 18) = 17.31$, $p < .0001$, $\hat{\eta}^2 = .233$].

First, from Figure 8, we can deduce that the *CP95* was too small when ML was used. Only when the number of participants was 7 and there were 40 observations, the ML is close to the nominal level. The *CP95* ranged from .618 to .948.

Second, the *CP95* was close to or larger than the nominal level of .95 when Bayesian procedures were applied (except from the inverse-Wishart prior distribution using the identity matrix). The nominal level of .95 was not attained in any condition when the number of participants was small (i.e., $J = 3$). When the number of participants was set to 5, all Bayesian procedures resulted in a *CP95* close to the nominal level (except from the inverse-Wishart prior distribution using the identity matrix). However, by further increasing the number of participants, the *CP95* became slightly too low (the *CP95* ranged from .87 to .90). These results can be found in Figure 8.

--- INSERT FIGURE 8 ABOUT HERE ---

Extensions to the Basic Two-Level Model Simulation Study: Informative Prior and Autocorrelation

A first extension to the basic two level model involves the construction of a more informative prior and its influence on parameter recovery. We will discuss this using the half-Normal distribution as functional form (but similar results were obtained for the half-Cauchy) and for a subset of design conditions: $\hat{\theta}_{00} = 0$, $\hat{\theta}_{10} = 0$ $\hat{\theta}_{20} = 2, \hat{\theta}_{30} = 0.2$, $J = 3$ or 5 or 7, $I = 20$, and $diag(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u3}) = diag(2, 0.2, 2, 0.2)$. The prior can be constructed based on the

same logic as discussed for the weakly informative prior specification for the half-Normal distribution, but now we include more information based on the re-analyses of meta-analyses (Table 1). As discussed before, we found an upper limit of 4.980 for the between-case variance of the immediate treatment effect. This corresponds to an upper standard deviation of 2.23 and as a consequence, the variance of the half-Normal corresponds to: 1.30 $(SD_u/1.96)^2$ [having a median of 1.08 (0.39 × 2.34)]. Therefore, we chose a variance of 2, which is informative, but would still not dominate the data and as a consequence we expect this prior resulting in better recovery of the variance components compared to the weakly informative priors. Here we compare the results of the informative prior with the weakly informative half-Normal distributions (having a variance of 10, 20 or 50).

Results. In terms of the relative bias and the mean squared error of the treatment effect estimates, similar results compared to the weakly informative priors are found: treatment effects are unbiased and the larger the number of cases, the more precise the estimate. For the *CP95*, we found differences between the analyses models [$F(3, 11)$] = 7.20, p = .02, $\hat{\eta}^2$ = .69], with the more informative prior resulting in a *CP95* closer to the nominal level across all conditions. The *CP95* equaled .94, .95 and .95 for 3, 4, and 7 cases respectively (whereas the *CP95* for the weakly informative priors (half-Normal with 10, 20 and 50 as variance) in the same conditions varied from .97 to 1.00).

For the variance components, the analysis procedure [$F(3, 5999)$] = 187.05, p < .001, $\hat{\eta}^2$ = .083] and the number of cases [$F(2, 5999)$] = 41.21, p < .001, $\hat{\eta}^2$ = .012] both have a statistical significant effect on the relative bias with the model having a larger effect, indicated by the larger value for $\hat{\eta}^2$. The between-case variance is underestimated for the informative prior, whereas it is overestimated for the weakly informative priors. The relative bias is larger than 5%

for all the conditions when using the informative prior, whereas unbiased estimates for the weakly informative priors are obtained when 7 cases are included. For the *MSE*, the analysis procedure [$F(3, 5999)$] = 187.05, $p < .001$, $\hat{\eta}^2 = .083$] and the number of cases [$F(2, 5999)$] = 41.21, $p < .001$, $\hat{\eta}^2 = .012$] have a statistical significant, but rather small effect: the more informative the prior, the smaller the *MSE* and the more cases included, the smaller the *MSE*. Although not statistically significant at the .001 level, the analysis procedure has only a moderate effect on the *CP95* [$F(3, 11)$] = 3.16, $p = .10$, $\hat{\eta}^2 = .18$], whereas the number of cases has a statistically significant and large effect on the *CP95* [$F(2, 11)$] = 18.43, $p = .003$, $\hat{\eta}^2 = .70$]. The *CP95* is overestimated in all conditions and all analyses models except when using the informative prior.

We investigated a second extension of the basic multilevel model by modeling autocorrelation. For this, a level-1 error covariance structure (Σ_e) was generated to follow the first-order autoregressive error structure, AR(1), with an autocorrelation value of .20. Shadish and Sullivan (2011) found that the distribution of AR values across 809 studies ranged from -.931 to .786 (which covers almost the whole theoretical range [-1,1]), with an average of .20. By adding autocorrelation to the model, we also have to specify a prior on the AR for the Bayesian procedure. Given the large range of possible values for the AR parameter, we used a uniform distribution with lower limit -1 and upper limit +1.

We found that the estimate of the AR is unbiased as the relative bias was smaller than or equal to 5% in all conditions (the value for the AR ranged from 0.19 to 0.21). No statistically significant effect of the design conditions on the relative bias or *MSE* was found. However, as the number of measurements and the number of participants increased, the smaller the relative

bias and precision. In addition, the presence of this AR did not influence the accuracy of the other parameter estimates.

Discussion

In current study, we described and evaluated the Bayesian estimation procedure as an alternative to the ML estimation procedure in contexts of two-level modeling of SCED data. We did not aim to convince the researcher of the Bayesian approach and philosophy, but rather discussed the use of Bayesian estimation procedures in contexts of SCED and elaborated on choosing and construction of prior distributions. Based on work by Gelman (2006) and Gelman et al. (2013) we proposed weakly informative priors using half-Cauchy, half-Normal distribution, and inverse-Wishart distributions for the Bayesian approach.

The results of the simulation study comparing the performance of ML (FML and REML) and Bayesian estimation procedures confirm the results of previous research about estimates of fixed effects (i.e., average immediate treatment effect and treatment effect on the slope) when using two-level modeling for multiple-baseline across participants designs (Ferron et al., 2009; Van den Noortgate & Onghena, 2003a, 2003b). The treatment effect estimates are unbiased (i.e., the relative bias is smaller than 5%) and precisely estimated. Current research confirms that REML is capable of producing *CP95*s close to the nominal level of .95 as is the case for the Bayesian procedures (except from the inverse-Wishart prior distributions) when there are at least five participants included. When there are only three participants included, the *CP95* is too small for the ML procedures and slightly too large for the Bayesian methods. As a consequence, the power of Bayesian procedures will probably be smaller compared to ML procedures. However, this was not formally evaluated in the simulation study. The ML does not attain the nominal confidence proportion of .95 in any condition. In sum, based on the results for the fixed effect

estimates, both REML and the Bayesian procedures seem appropriate when the number of participants is equal to or larger than five.

An asset of the current study compared to previous research is that we also evaluate the variance components estimates. The between-participant variance estimates of the treatment effects yield important information as a means to evaluate whether the treatment has a similar effect across the participants or whether there is a large amount of variability. Previous research using REML indicates it produces biased estimates of the variance components (Ferron et al., 2009). When the number of participants is small (i.e., $J = 3$), biased and less precise estimates are found for both the ML and Bayesian methods. When more participants are included in the synthesis (i.e., $J = 5$) less biased and more precise estimates are obtained and the difference in estimates between the procedures diminishes (except when the inverse-Wishart with identity matrix is used). The *CP95* of the between-participant standard deviation is dependent on the analysis method. Even with a small number of participants, the *CP95* is close to the nominal level of .95 for the Bayesian procedures (except when the inverse-Wishart is used) in contrast to the ML procedures. For the ML procedures, at least seven participants having 40 measurements are needed. Previous research recommended using REML in cases with small sample sizes. However, we want to warn SCED data analysts to not solely rely on REML if the research interest lies in the variance components estimates. In addition, we explored the option of constructing a more informative prior. The more informative prior resulted in more precise estimates and *CP95* for the fixed effects and the random components closer to the nominal level. However, still biased variance components were obtained. The construction of informative priors and exploring different functional forms and its influence on the parameter recovery are beyond the scope of this study and we recommend this as a first step for future research in this area.

In sum, when the researcher wants to estimate the two-level model parameters using Bayesian methods, the half-Normal, and the half-Cauchy distributions can be used as priors, at least in conditions similar to the ones examined in this study and when at least five participants were included. If only three participants were included, all procedures resulted in biased estimates, but more precise estimates are obtained when using maximum likelihood and half-Normal prior distributions. In addition, in these conditions, the *CP95* for the Bayesian procedures is close to the nominal level in contrast to the ML procedures. The inverse Wishart with the identity matrix seems less appropriate in terms of bias of the parameter estimates and *CP95*. By using this standard noninformative version of the inverse-Wishart prior (small df and identity matrix) the marginal distribution of the correlations is uniform. Large standard deviations are related with large absolute correlations, which is not non-informative. In order to deal with this, the variances can be estimated first, and then use this to tweak the inverse-Wishart prior to have the right scale (Kass & Natarajan, 2006). This is what we did using the non-identity matrix.

The construction of confidence intervals around the estimates of the variance components seemed to be less straightforward as the estimated variance components distribution is extremely skewed. SAS Proc MIXED uses a Satterthwaite approximation that seems to results in unrealistic large upper limits of the confidence interval in certain conditions. In this study, we proposed an adjustment to the formula standard programmed in SAS 9.4 that resulted in more reasonable values. However, further research in constructing confidence intervals around variance estimates is needed.

Extensions

As with any simulation study, the results are limited to the chosen conditions and cannot be generalized to other conditions. Although we included commonly encountered conditions for SCED research, we are aware that the data and/or research questions may require more complicated models making our simulation results less informative. In this situation, the researcher can replicate the simulation study and slightly change the values depending on the prior belief and conduct a sensitivity analysis. In addition, we believe that further systematic simulation research is needed on the performance of ML and Bayesian procedures for SCED data in extensions of the studied conditions, as described as follows.

Non-Continuous Outcomes. In a review of the SCED literature, Shadish and Sullivan (2011) indicated that over 90% of the outcome measures used were some form of count. These counts can be a number out of a fixed total, a number without a total, or can be presented as a percentage or rate. This is confirmed by a recent study aimed at giving an overview of SCED data characteristics by coding data for 399 SCEDs that appeared in the Journal of Applied Behavior Analysis in 2012 (Ugille et al., 2015). Ugille et al. (2015) found that the majority of SCED outcomes (i.e., 77.19%) are counts of some sort. Variables of these types are often poorly approximated by a Gaussian error distribution. For instance, floor or ceiling effect in one of the phases of an experiment or substantial heteroscedasticity across phases are not unusual. Failing to account for these features can end up creating important biases in quantitative effect size measures derived from the model. For instance, effect sizes are often extremely large and out of a plausible range.

In scenarios where there are no floor or ceiling effects (i.e., percentages near zero or 100 percent), percentages can be treated as continuous outcomes (Kratochwill & Levin, 2014) and as a consequence similar priors suggested as for continuous outcomes can be used as proposed in

this study. In the other scenarios (only representing a small fraction of the SCED data), in which the outcome is a count or rate, or percentages with floor and ceiling effects, it is better to model the count outcomes as counts by using a Poisson distribution (in the case the outcome is a count during an interval) or a binomial distribution (i.e., percentages or rates). In these situations future research is needed to define priors.

Non-linear Trajectories. Trajectories in MBD data may be non-linear (i.e., Beretvas, 2011, Hembry, Bunuan, Beretvas, Ferron & Van den Noortgate, 2014, Mulloy, 2011, Shadish et al., 2013). For instance, the effects can be increasing quadratically or logistically. Indeed, a typical MBD data trajectory is characterize by upper and/or lower asymptotes and can therefore be best described by a logistic model as suggested by Beretvas (2010) and validated through a consecutive simulation study of Hembry et al. (2014). In their study, Bayesian procedures were used, defining non-informative priors on the fixed effect, an inverse gamma on the within-participant residual variance and half-Cauchy distributions on the between-participant standard deviation.

Other Type of SCEDs. In this study, we focused on MBDs across participants because they represent the majority of the published SCED designs (i.e., 54,3% of the 809 SCEDs coded by Shadish & Sullivan, 2011). However, there are a variety of other type of SCEDs such as alternating treatment designs and reversal designs. Previous research already suggested coding of these design matrices (i.e., Shadish et al., 2013; Moeyaert, et al., 2014).

Covariance Structures. In line with previous simulation research on multilevel modeling of small-*N* studies, we simulated and analyzed data assuming no covariance between the regression coefficients, resulting in a diagonal covariance matrix at the second level (Baldwin & Fellingham, 2013; Shadish, Hedges, Pustejovsky, Rindskopf, Boyajian, & Sullivan, 2014,).

However, if one is interested in analyzing the dataset with covariances, the inverse-Wishart distribution could be used as suggested by Gelman et al. (2013). However, we did not explore this in current study as the inverse-Wishart distribution needs further research as biased and less precise estimates can be obtained. Previous research warns that when variances are small (as is the case in SCED contexts) the inverse-Wishart prior specification can have a considerable impact on the parameter estimates.

Also within participants, more complex covariance structures can be studied, such as higher order autoregressive models and first-order moving averages (Baek & Ferron, 2013). We also assumed homogeneous within-participant variance whereas the variance in outcome scores might be smaller in the baseline phase compared to the treatment phase and as such, heterogeneous within-participant variance might be more reasonable.

Other prior distributions. We used half-Normal distributions, half-Cauchy distributions, and inverse-Wishart as priors, but other distributions are also possible such as the *t*-distribution, the uniform distribution etc.. We invite other researchers to build further on this explorative simulation study, by varying the parameter values of the priors and the type of priors. A limitation of our simulation study is that we cannot determine when a chosen prior distribution has a too large effect on the results of the analysis. In a future study, it would be interesting to focus on the conditions in which three, five and seven participants are included (keeping all other conditions and parameter values constant), construct informative priors on the between-participant standard deviation and systematically vary the degree of information included in the informative prior in order to investigate when results are overly influenced by the chosen prior distribution.

Additional levels. Another next step in this research field might be to add an additional level to the multilevel model, representing the study level. Recently there has been an increase in published meta-analyses of SCEDs. Meta-analyses investigating the same research question could be combined (to get more general treatment effect estimates) and this implies three-level modeling (measurements are clustered within participants and participants in turn are clustered within studies).

Conclusion

Currently, Bayesian techniques are not commonly taught in introductory statistics classes, and misconceptions exist concerning adding prior knowledge to a model. Defining priors can be hard, but seems natural to account for information that is already available in the existing literature before drawing conclusions in a study. In order to relax asymptotic assumptions (needed in the frequentist framework) and in order to make immediate probability statements, a researcher might lean towards Bayesian estimation procedures from a conceptual point of view.

The Bayesian models result in similar results compared to the ML if the number of participants is at least five. However, if the number of participants is three, biased and less precise estimates are obtained for both approaches, but the *CP95* is closer to the nominal level using the Bayesian approach (except from the inverse-Wishart distribution). The inverse-Wishart prior distribution with the identity matrix is not recommended for the conditions investigated in this simulation study. If a researcher applies the Bayesian data analysis approach to real data, we recommend to use a variety of different prior distributions and discuss to what extent and in what sense the results depend on the prior used. The more informative prior resulted in more precise estimates and *CP95* for the fixed effects and the random components closer to the nominal level.

References

Alen, E., Grietens, H., & Van den Noortgate, W. (2009). *Meta-analysis of single-case studies: An illustration for the treatment of anxiety disorders.* Unpublished manuscript.

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45,* 65-74.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97. doi: 10.1901/jaba.1968.1-91

Baldwin, S. A. & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*, 151-164.

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., … Watson, J. (2011). Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. *Cognitive Behavior Therapy, 40*, 15-33. Doi: 10.1080/16506073.2010.520731

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn & Bacon.

Beretvas (2011). *A multilevel model for nonlinear trajectories with smooth transitions for multiple baseline design data.* Paper presented at the annual meeting of the American Educational Association, New Orleans, LA.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized liner mixed models. *Journal of American Statistical Assosiation, 88,* 9-24.

Browne, W. J., Draper, D., Goldstein, H. and Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis*, *39*, 203-225.

Burrick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components*. New York, NY: Marcel Dekker.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cohen, J. (Ed.). (1988). *Statistical power analysis for the behavioural sciences.* United states of America: Lawrence Erlbauw Associates.

Daniels, M. J., & Kass, R. E. (2001), Shrinkage estimators for covariance matrices. *Biometrics*, 57, 1173-1184.

Depaoli, S. (2014). The impact of inaccurate informative priors for growth parameters in Bayesian Growth Mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12,* 239-252. Doi: 10.1080/10705511.2014.882686

Depaoli, S. (2015). Mixture class recovery in GMM under verying degrees of class separation: frequentist versus Bayesian estimation, *Psychological Methods, 18,* 186-219. Doi: 10.1037/a0031609

DerSimonian, R. (1996). Meta-analysis in the design and monitoring of clinical trials. Statistics in Medicine, 15, 1237-1248.

Denis, J., Van den Noortgate, W., & Maes, B. (2011). Self-injurious behavior in people with profound intellectual disabilities: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 32*, 911-923. doi: 10.1016/j.ridd.2011.01.014

de Vries, R. M. & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*, 165-185. doi: 10.1037/a0031037.

Farmer, J., Owens, C. M., Ferron, J. M., & Allsopp, D. (2010). *A review of social science single-case meta-analyses*. Manuscript in preparation.

Ferron, J. M., Bell, B. A., Hess, M. F., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: the utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372-384. doi: 10.3758/BRM.41.2.372

Ferron, J. M., & Scott, H. (2005). Multiple baseline designs. In B. Everitt & D. Howell (Eds). *Encyclopedia of Behavioral Statistics* (Vol. 3, pp. 1306-1309). West Sussex, UK: Wiley & Sons Ltd.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B.(2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Hill, L. S. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.

Goldstein, H. (1995). *Multilevel statistical models*. London, England: Edward Arnold.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72* (358), 320-340. doi: 10.2307/2286796

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109. doi: 10.2307/2334940

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2014). Estimation of a nonlinear intervention phase trajectory for multiple-baseline data. *Journal of Experimental Education, 83,* 514–546. doi: 10.1080/00220973.2014.907231

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367. doi: 10.1177/0049124198026003003

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Erlbaum.

Huitema, B. E., & J. W. McKean. (1994). Two reduced-bias autocorrelation estimators: rF1 and rF2. *Perceptual and Motor Skills,* 78, 323–330.

Jackman, S. (2009). *Bayesian analysis for social sciences.* New York, NY: Wiley. doi: 10.1002/9780470686621

Kass , R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association, 91*, 1343-1370.

Kazdin, A.E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from Restricted Maximum Likelihood. *Biometrics*, 53, 983–997. doi: 10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from Restricted Maximum Likelihood. *Computational Statistics and Data Analysis*, 53, 2583–2595. doi: 10.1016/j.csda.2008.12.013

Kinugasa, T., Cerin, E., & Hooper, S. (2004). Single-subject research designs and data analyses for assessing elite athletes' conditioning. *Sports Medicine, 34*, 1035-1050. doi: 10.2165/00007256-200434150-00003

Koehler, M. J., & Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments & Computers, 32*, 367-371. doi: 10.3758/BF03207807

Kokina, A., & Kern, L. (2010). Social story interventions for students with autism spectrum disorders: a meta-analysis. *Journal of Autism and Developmental Disorders, 40*, 812-826. doi: 10.1007/s10803-009-0931-0

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124–144. doi:10.1037/a0017736

Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

Lau, J., Schmid, C. H. & Chalmers, T. C. (1995). Cummulative meta-analysis of clinical trials builds evidence for exemplary medical care. Journal of Clinical Epidemology, 48, 45-57.

Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological intervention research. In I. B. Weiner (Series Ed.) and W. M. Reynolds & G. E. Miller (Vol. Eds.). *Handbook of psychology: Vol. 7. Educational psychology* (pp. 557–581). Hoboken, NY: Wiley.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientistis.* New York, NY: Springer. doi: 10.1007/978-0-387-71265-9

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*, 201-215.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychol Methods, 5* (1), 87-101.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092. doi: 10.1063/1.1699114

Moeyaert, M., Maggin, D. M., & Verkuilen, J. (2016). Reliability and validity of extracting data from image files in contexts of single-case experimental design studies. *Behavior Modification.* Advance online publication. doi: 10.1177/0145445516645763

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, T., & Van den Noortgate, W. (2013a). Modeling external events in the three-level analysis of multiple-baseline across participants designs: A simulation study. *Behavior Research Methods, 45,* 547-559. doi: 10.3758/s13428-012-0274-1

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, N., & Van den Noortgate, W. (2013b). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education, 82,* 1-21. doi: 10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2013c). The three-level synthesis of standardized single-subject experimental data: a Monte Carlo simulation study. *Multivariate Behavioral Research, 48*, 719-748. doi: 10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of

single-subject experimental design research. *Behavior Modification, 38* (5), 665-704. doi: 10.1177/0145445514535243

Mulloy, A. M. (2011). A Monte Carlo investigation of multilevel modeling in meta-analysis of single-subject research data. (Doctoral dissertation).

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850-1854). Chichester: Wiley.

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44,* 795-805. doi: 10.3758/s13428-011-0180-y

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*, 545-554. doi: 10.2307/2334389

Pauler, D.K. and Wakefield, J.C. (2000). Issues in the modeling and implementation of Bayesian meta-analyses. *Meta-Analysis in Medicine and Health Policy*, Stangl, D. and Berry, D.A. (Eds), 205-230. Marcel Dekker, New York and Basel.

Rantz, W. G., Dickinson, A. M., Sinclair, G. A., & Van Houten, R. (2009). The effect of feedback on the accuracy of checklist completion during instrument flight training. *Journal of Applied Behavior Analysis, 42*, 497-509.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods. Second edition* (Vol. 1). London, New Delhi.

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

Robinson, G. K. (1991). That BLUP is a good-thing: The estimation of random effects. *Statistical Science, 6*, 15-51.

Rohatgi, A. (2014). *WebPlotDigitizer user manual version 3.4*. Retrieved from http://arohatgi.info/WebPlotDigitizer/userManual.pdf

Shadish, W.R., Hedges, L. V., Pustejovsky, J., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2014). Analyzing single-case designs: d, G, multilevel models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analysis. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances.* Washington, DC: American Psychological Association.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological Methods, 18,* 385-405.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*, 95-109. doi: 10.1002/ev.217

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

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

Snijders, T., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* Hoboken, NK: Wiley.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., & Lunn, D. (1994, 2003). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England,http://www.mrc-bsu.cam.ac.uk/bugs/.

Thompson, R. (1980). Maximum likelihood estimation of variance components. *Mathematische Operationsforschung und Statistik, Series Statistics, 11,* 545-561.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. *Columbia Univ., New York, NY, USA, Tech. Rep*.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). *Characteristics of single-case experimental designs.* Manuscript submitted for publication.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative intergration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 34*, 1-10.

Van den Noortgate, W., & Onghena, P. (2003c). Mulitlevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*, 765-790.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention, 2*, 142-151.

Verkuilen, J., & Moeyaert, M. (2016). *A quasi-likelihood generalized estimating equation approach to rates and counts outcomes in single case experimental design.* Manuscript in preparation.

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: a meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5*, 562-569. doi: 10.1016/j.rasd.2010.06.023

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31,* 374-383.

Table 1

Parameter Estimated	Estimate (*SE*)

Overview of Re-analyses of Published Meta-Analysis of Single-Case Experimental Data

	Study 1	Study 2	Study 3	Study 4	Study 5
Fixed Effects					
Average baseline level (θ_{00})	6.593 (1.236)	3.227 (0.975)	4.160 (0.912)	4.201 (0.898)	1.087 (0.398)
Average trend during baseline (θ_{10})	-0.228 (0.095)	0.090 (0.076)	0.007 (0.012)	0.142 (0.063)	-0.007 (0.006)
Average treatment effect (θ_{20})	-0.909 (0.359)	-3.096 (0.756)	-1.133 (0.925)	-4.093 (0.987)	1.032 (0.256)
Average treatment effect on time trend (θ_{30})	-0.052 (0.007)	-0.113 (0.081)	-0.015 (0.044)	-0.165 (0.074)	0.015 (0.019)
Random Effects					
Between-study variance					
Between-study variance baseline level ($\sigma^2_{v_0}$)	12.021 (7.478)	11.671 (5.685)	8.621 (4.400)	0.927 (3.660)	0.888 (0.585)
Between-study variance trend baseline ($\sigma^2_{v_1}$)	0.093 (0.046)	0.075 (0.034)	0.000 (/)	0.018 (0.013)	0.000 (/)
Between-study variance treatment ($\sigma^2_{v_2}$)	0.329 (0.474)	7.156 (3.349)	10.708 (4.623)	1.401 (4.391)	0.190 (0.217)
Between-study variance treatment effect on trend ($\sigma^2_{v_3}$)	0.000 (/)	0.081 (0.0386)	0.015 (0.010)	0.011 (0.016)	0.001 (0.001)
Between-case variance					
Between- case variance baseline level ($\sigma^2_{u_0}$)	6.934 (2.544)	4.619 (2.031)	7.957 (2.355)	3.699 (5.276)	0.007 (0.033)
Between- case variance trend baseline ($\sigma^2_{u_1}$)	0.00002 (0.00005)	0.001 (0.001)	0.000 (/)	0.000 (/)	0.00005 (0.00008)
Between- case variance treatment ($\sigma^2_{u_2}$)	1.543 (0.572)	2.178 (1.043)	3.752 (1.217)	4.980 (5.892)	0.268 (0.156)
Between- case variance treatment effect on trend ($\sigma^2_{u_3}$)	0.0001 (0.00007)	0.002 (0.002)	0.002 (0.002)	0.000 (/)	0.000 (/)
Within-case residual variance	1.052 (0.033)	1.063 (0.042)	1.078 (0.061)	1.146 (0.110)	1.000 (0.058)

Note. Study 1 refers to the study of Alen et al. (2009); Study 2 Denis et al. (2008), Study 3 to Kokina and Kern, 2010; Study 4 to Shogren et al. (2004); and Study 5 to Wang et al. (2001). For completeness, we included the between-study variance estimates, but for this study we are mainly interested in the between-case variance estimates.

Table 2

Overview of the Prior Distributions for the Variance Components

Prior Distribution	Scenario 1	Scenario 2	Scenario 3
Half-Cauchy	Half-Cauchy ~ (0, 50)	Half-Cauchy ~ (0, 20)	Half-Cauchy ~ (0, 10)
Half-Normal	Half-Normal ~ (0, 50)	Half-Normal ~ (0, 20)	Half-Normal ~ (0, 10)
Inverse-Wishart	$iwish\begin{bmatrix}1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1\end{bmatrix}, 6$	$iwish\begin{bmatrix}2 & & & \\ 0 & 0.2 & & \\ 0 & 0 & 2 & \\ 0 & 0 & 0 & 0.2\end{bmatrix}, 6$	

Table 3

Start Points of the Treatment across Participants Within a Multiple-Baseline Design

J		*I* = 20	*I* = 40
3	Participant 1	7	11
	Participant 2	11	21
	Participant 3	15	31
7	Participant 1	7	11
	Participant 2	9	15
	Participant 3	9	15
	Participant 4	11	21
	Participant 5	13	27
	Participant 6	13	27
	Participant 7	15	31

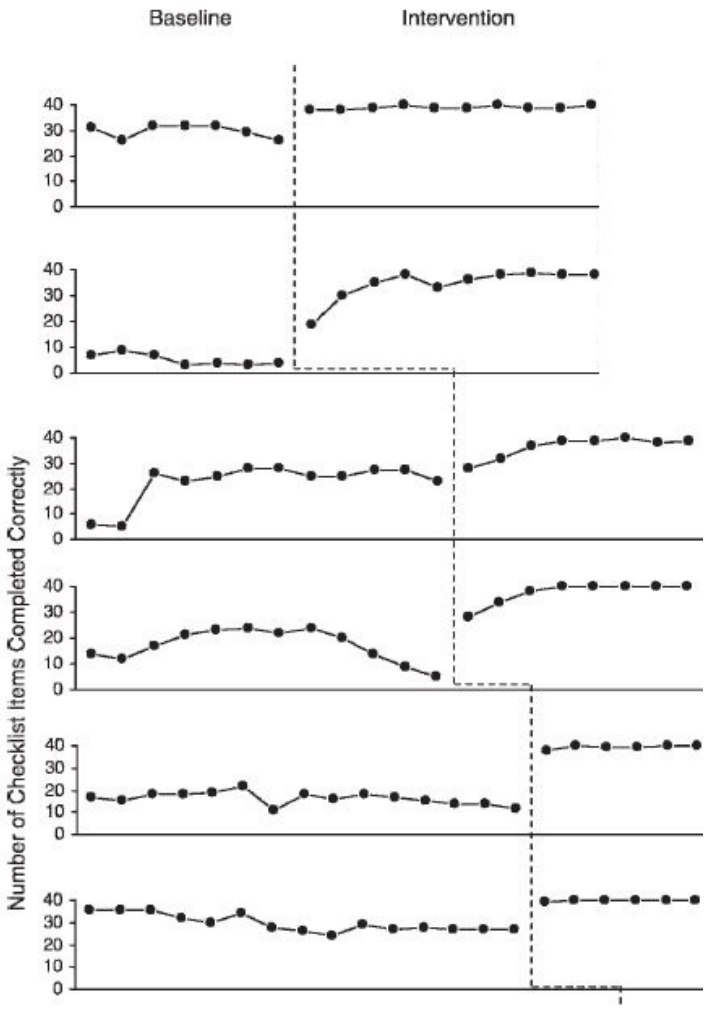

Figure 1. Graphical display of an MBD across 6 participants. The figure is adapted from 'The Effect of Feedback on the Accuracy of Checklist Completion during Instrument Flight Training", by W. G. Rantz, A. M. Dickinson, G. A. Sinclair, and R. V. Van Houten, 2009, *Journal of Applied Behavior Analysis, 42*, p. 503.

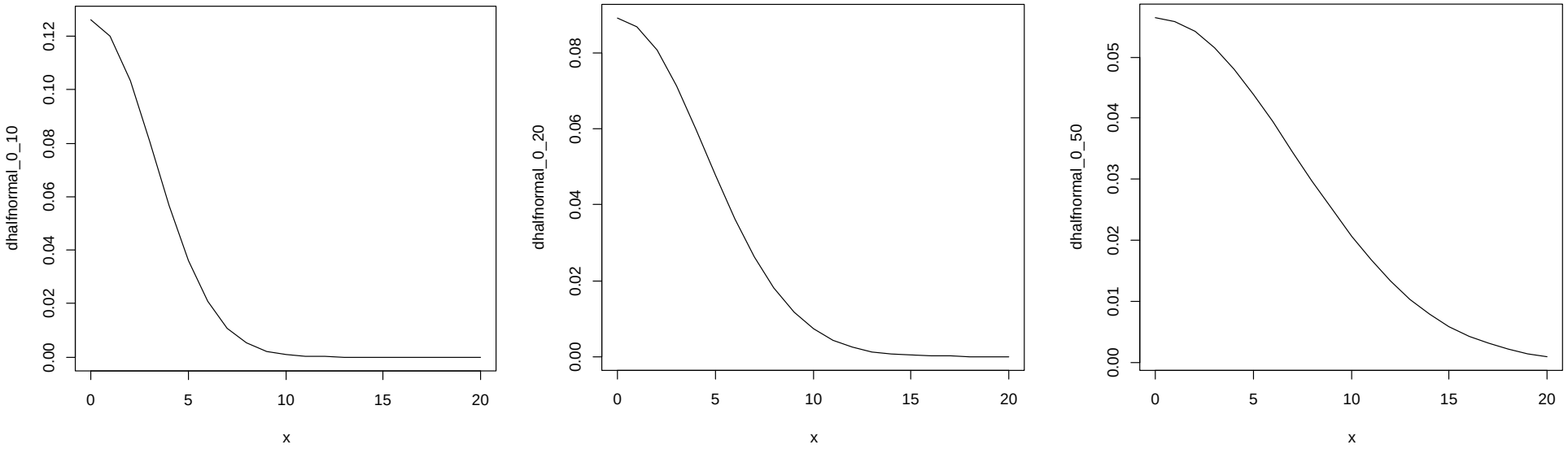

Figure 2. Illustration of the influence of the variance of the distribution on the shape of the half-Normal prior distribution. The left panel displays $X \sim N(0, 10,\ lower = 0.00)$, the middle panel displays $X \sim N(0, 20,\ lower = 0.00)$ and the right panel displays $X \sim N(0, 50,\ lower = 0.00)$.

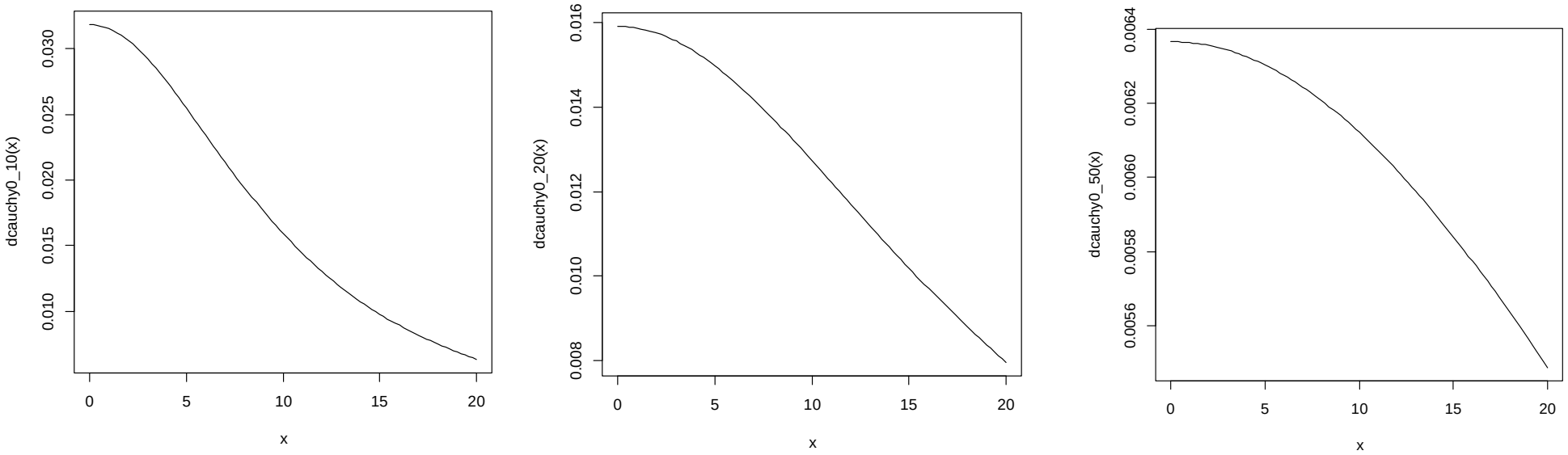

Figure 3. Illustration of the influence of the scale parameter on the form of the half-Cauchy prior distribution. The left panel displays $X \sim Cauchy(0, 10,\ lower = 0.00)$, the middle panel displays $X \sim Cauchy(0, 20,\ lower = 0.00)$ and the right panel displays $X \sim Cauchy(0, 50,\ lower = 0.00)$.

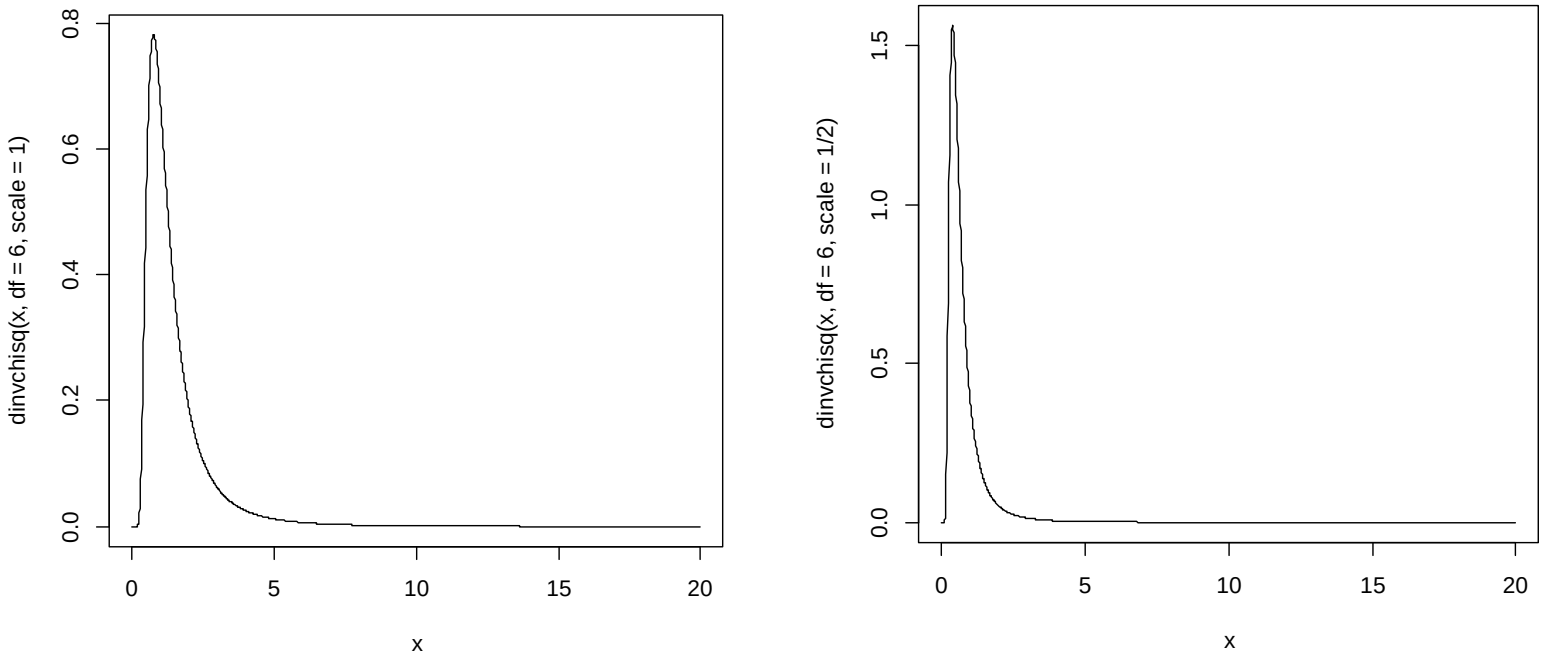

Figure 4. Illustration of the influence of the scale parameter on the form of the scaled inverse χ^2 distribution used as simplifications of the inverse Wishart distribution. The left panel displays $X \sim Inv-\chi^2(6,\ 1)$, representing the simplified version of the inverse Wishart distribution using the identity matrix and the right panel displays $X \sim Inv-\chi^2(6,\ 1/2)$ representing the simplified version of the inverse Wishart distribution using the non-identity matrix (assuming a variance of 2).

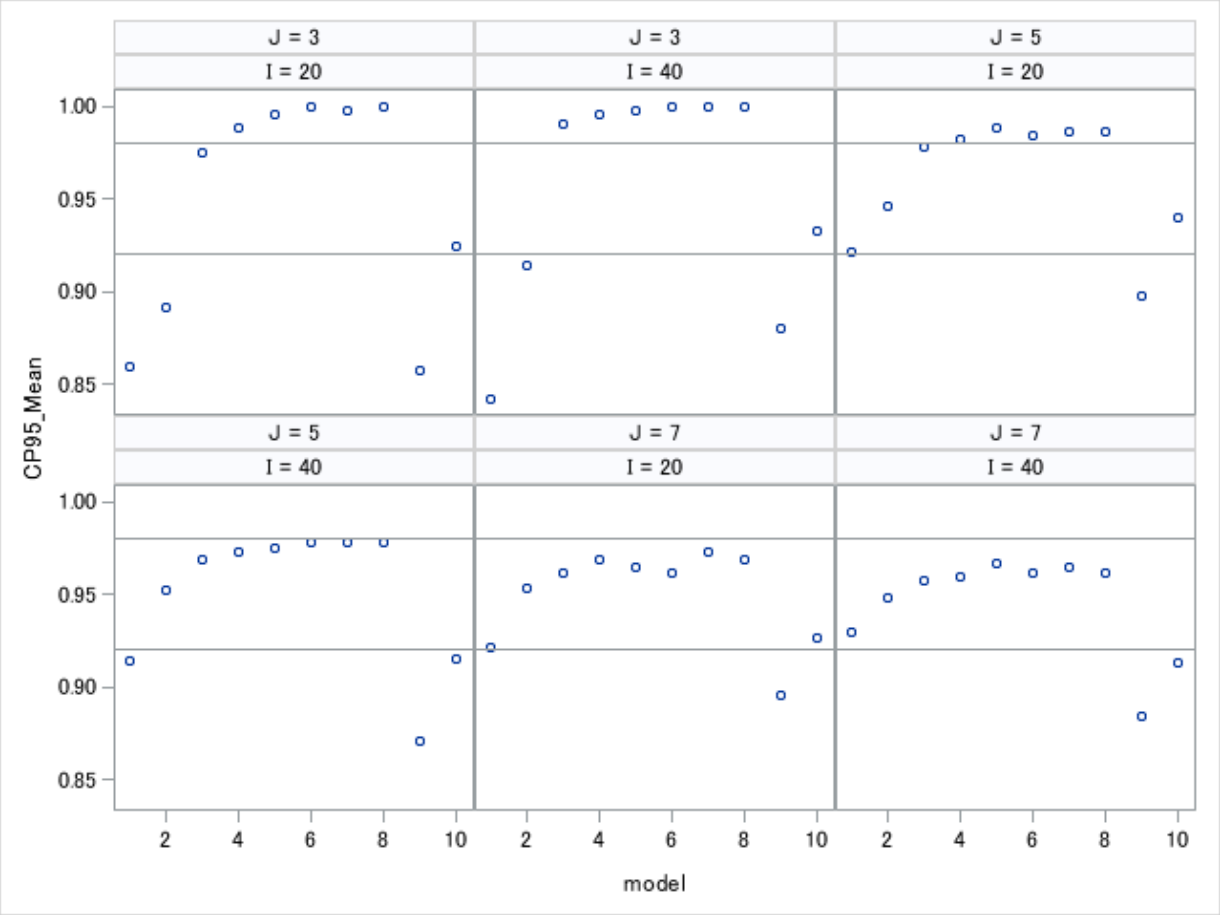

Figure 5. Graphical display of the coverage proportion of the 95% confidence interval of the immediate treatment effect estimate as a function of the number of participants, measurements and the analysis procedure. Full information Maximum Likelihood = Model 1, Restricted Maximum Likelihood = Model 2, Half-Normal ~ (0, 10) = Model 3, Half-Normal ~ (0, 20) = Model 4, Half-Normal ~ (0, 50) = Model 5, Half-Cauchy ~ (0, 10) = Model 6, Half-Cauchy ~ (0, 20) = Model 7, Half-Cauchy ~ (0, 50) = Model 8, Inverse Wishart with identity matrix = Model 9 and with inverse Wishart with non-identity values = Model 10. CP95 = coverage proportion of the 95% confidence interval. *I* and *J* indicate the number of measurements and cases respectively.

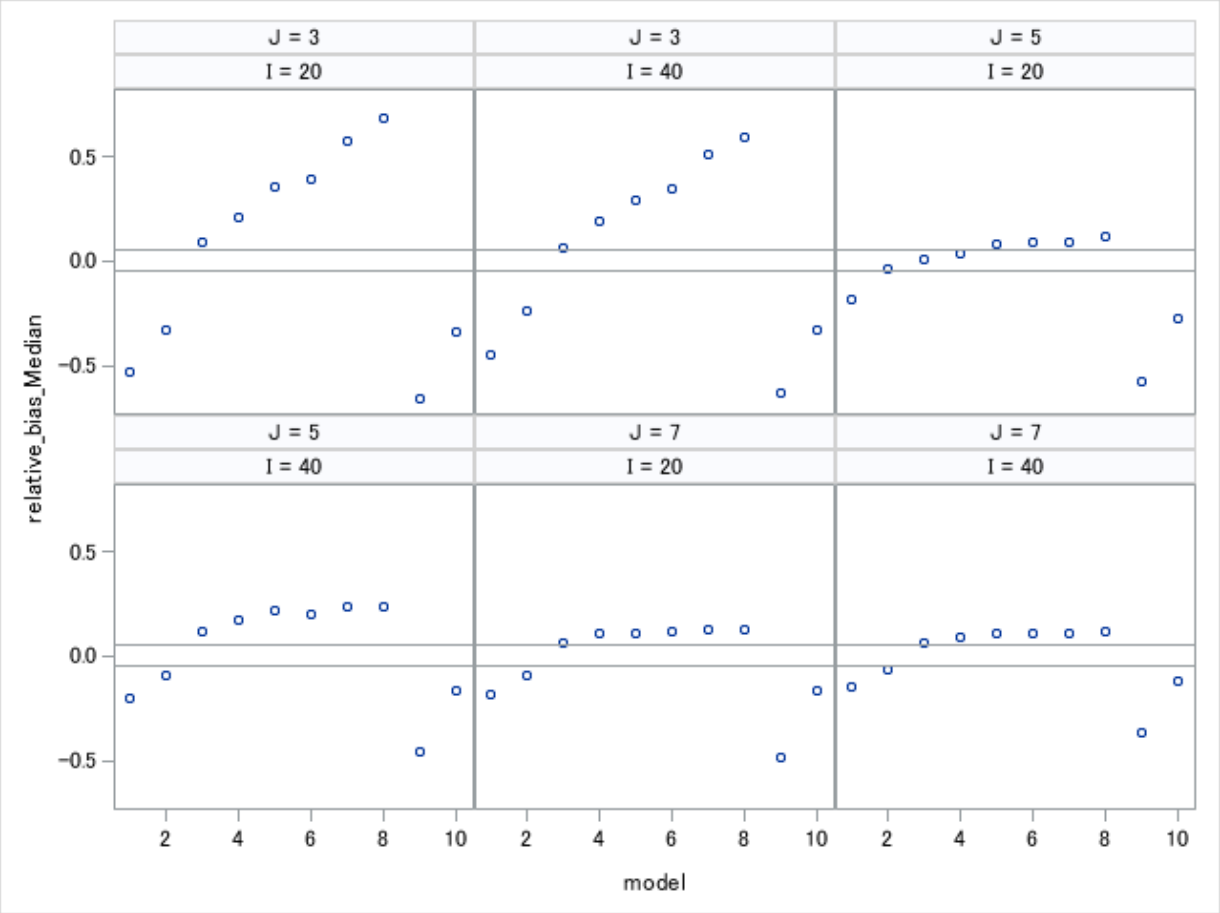

Figure 6. Graphical display of the relative bias of the between-case standard deviation estimate as a function of the number of participants, measurements and the analysis procedure.
Full information Maximum Likelihood = Model 1, Restricted Maximum Likelihood = Model 2, Half-Normal ~ (0, 10) = Model 3, Half-Normal ~ (0, 20) = Model 4, Half-Normal ~ (0, 50) = Model 5, Half-Cauchy ~ (0, 10) = Model 6, Half-Cauchy ~ (0, 20) = Model 7, Half-Cauchy ~ (0, 50) = Model 8, Inverse Wishart with identity matrix = Model 9 and with inverse Wishart with non-identity values = Model 10. *I* and *J* indicate the number of measurements and cases respectively.

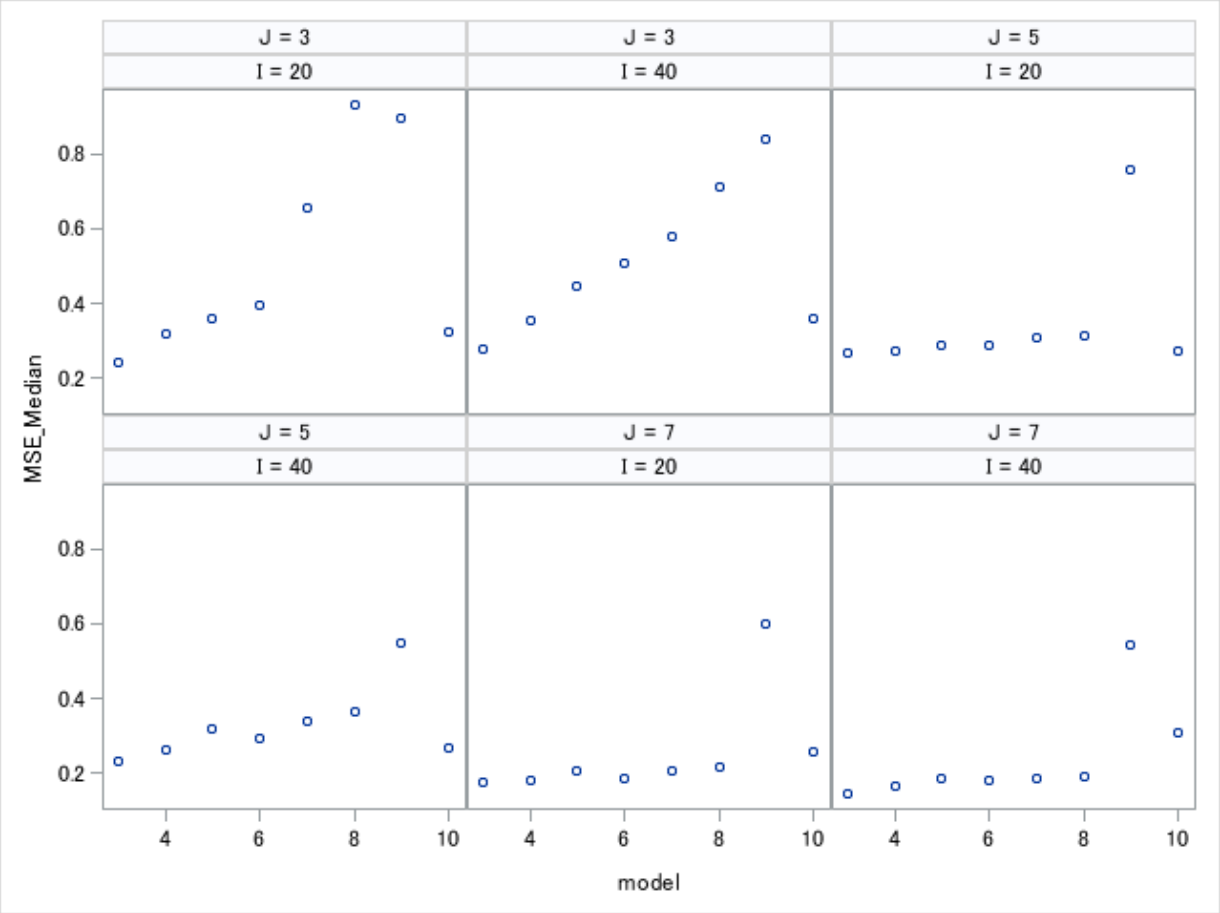

Figure 7. Graphical display of the mean squared error of the between-case standard deviation estimate as a function of the number of participants, measurements and the analysis procedure. Full information Maximum Likelihood = Model 1, Restricted Maximum Likelihood = Model 2, Half-Normal ~ (0, 10) = Model 3, Half-Normal ~ (0, 20) = Model 4, Half-Normal ~ (0, 50) = Model 5, Half-Cauchy ~ (0, 10) = Model 6, Half-Cauchy ~ (0, 20) = Model 7, Half-Cauchy ~ (0, 50) = Model 8, Inverse Wishart with identity matrix = Model 9 and with inverse Wishart with non-identity values = Model 10. MSE = Mean Squared Error. *I* and *J* indicate the number of measurements and cases respectively.

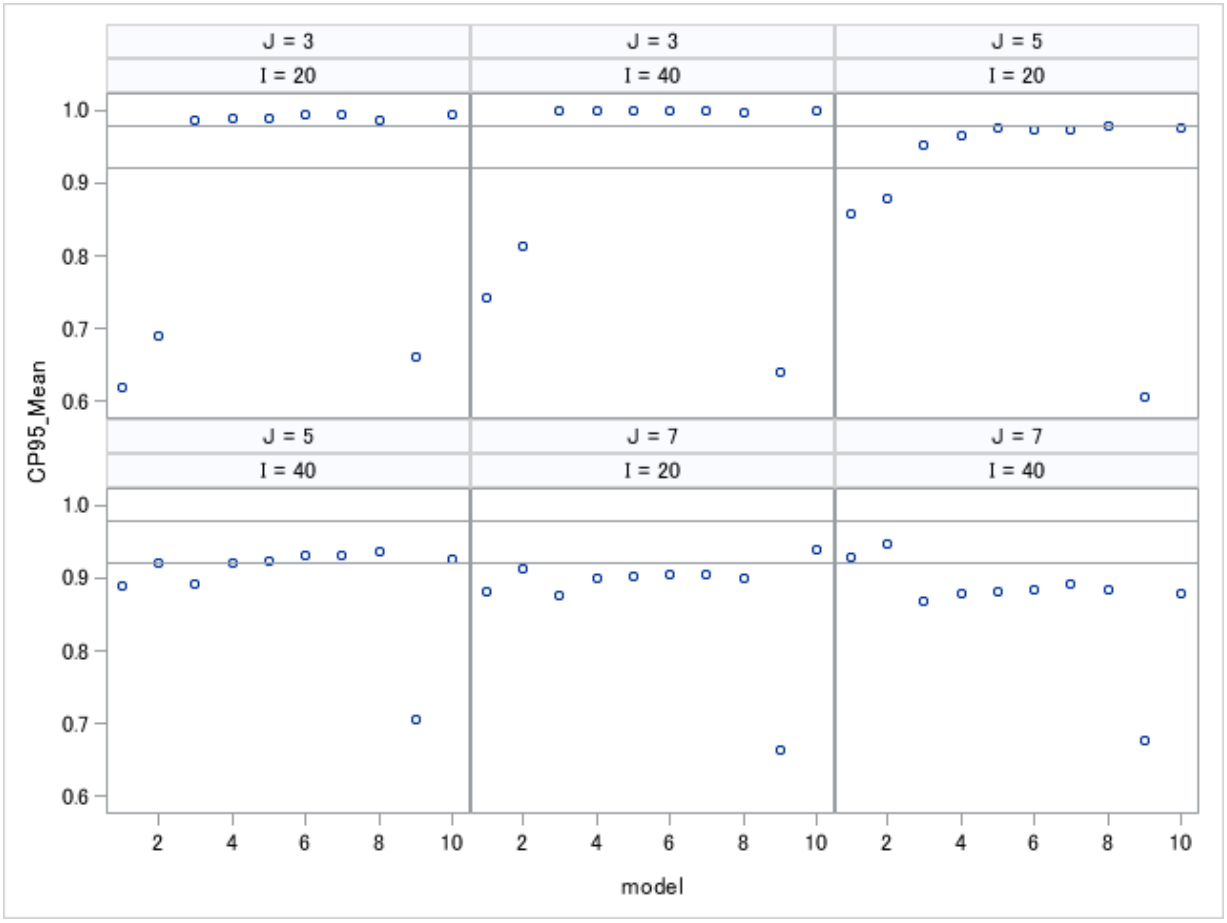

Figure 8. Graphical display of the coverage proportion of the 95% confidence interval of the between-case standard deviation estimate as a function of the number of participants, measurements and the analysis procedure. Full information Maximum Likelihood = Model 1, Restricted Maximum Likelihood = Model 2, Half-Normal ~ (0, 10) = Model 3, Half-Normal ~ (0, 20) = Model 4, Half-Normal ~ (0, 50) = Model 5, Half-Cauchy ~ (0, 10) = Model 6, Half-Cauchy ~ (0, 20) = Model 7, Half-Cauchy ~ (0, 50) = Model 8, Inverse Wishart with identity matrix = Model 9 and with inverse Wishart non-identity values = Model 10. *CP95* = coverage proportion of the 95% confidence interval. *I* and *J* indicate the number of measurements and cases respectively.