

Modeling Clustered Data with Very Few Clusters

Daniel McNeish & Laura M. Stapleton

To cite this article: Daniel McNeish & Laura M. Stapleton (2016) Modeling Clustered Data with Very Few Clusters, *Multivariate Behavioral Research*, 51:4, 495-518, DOI: [10.1080/00273171.2016.1167008](https://doi.org/10.1080/00273171.2016.1167008)

To link to this article: <https://doi.org/10.1080/00273171.2016.1167008>



© 2016 The Author(s). Published with license by Taylor & Francis© Daniel McNeish, and Laura M. Stapleton.



[View supplementary material](#)



Published online: 07 Jun 2016.



[Submit your article to this journal](#)



Article views: 13186



[View related articles](#)



[View Crossmark data](#)



Citing articles: 87 [View citing articles](#)

Modeling Clustered Data with Very Few Clusters

Daniel McNeish^{a,b} and Laura M. Stapleton^a

^aUniversity of Maryland; ^bUtrecht University

ABSTRACT

Small-sample inference with clustered data has received increased attention recently in the methodological literature, with several simulation studies being presented on the small-sample behavior of many methods. However, nearly all previous studies focus on a single class of methods (e.g., only multilevel models, only corrections to sandwich estimators), and the differential performance of various methods that can be implemented to accommodate clustered data with very few clusters is largely unknown, potentially due to the rigid disciplinary preferences. Furthermore, a majority of these studies focus on scenarios with 15 or more clusters and feature unrealistically simple data-generation models with very few predictors. This article, motivated by an applied educational psychology cluster randomized trial, presents a simulation study that simultaneously addresses the extreme small sample and differential performance (estimation bias, Type I error rates, and relative power) of 12 methods to account for clustered data with a model that features a more realistic number of predictors. The motivating data are then modeled with each method, and results are compared. Results show that generalized estimating equations perform poorly; the choice of Bayesian prior distributions affects performance; and fixed effect models perform quite well. Limitations and implications for applications are also discussed.

KEYWORDS

Bayesian; cluster randomized trial; fixed effect model; GEE; HLM; multilevel model; small sample


Clustered data with few clusters are quite common in behavioral sciences due to practical concerns such as financial limitations, the use of extant data sets, or difficulties in recruiting large numbers of participants. For example, it is expensive to recruit many higher-level units such as schools or hospitals to participate in a research study; secondary data sets may include survey information based on a limited population; and certain populations may simply be sparsely distributed and not large, making it challenging to gather a large sample (e.g., schools specifically for deaf students in the United States).

Over the past decade, several simulation studies have addressed the small-sample properties of a variety of methods for clustered data including both multilevel models (MLMs; e.g., Bell, Morgan, Schoenberger, Kromrey, & Ferron, 2014; Browne & Draper, 2006; Hox, van de Schoot, & Matthijsse, 2012; Maas & Hox, 2004; 2005) and so-called design-based methods¹ such as generalized estimating equations (GEE) or cluster-robust errors (Angrist & Pischke, 2008; Cameron, Gelbach, & Miller, 2011; Emrich & Piedmonte, 1992; Gunsolley, Gerschell, & Chinchilli, 1995; Lu et al., 2007; Morel, Bokossa, & Neerchal, 2003; Pan & Wall, 2002;

Westgate, 2013). Very broadly, these studies generally show that models with about 20 to 40 clusters exhibit desirable properties (e.g., consistency) and that certain small-sample corrections such as the Kenward-Roger correction (Kenward & Roger, 1997; 2009) for MLMs and the Mancl-DeRouen (Mancl & DeRouen, 2001), Kauermann-Carroll (Kauermann & Carroll, 2001), Morel-Bokossa-Neerchal (Morel et al., 2003), and Fay-Graubard (Fay & Graubard, 2001) corrections for GEE were able to maintain desirable statistical properties with as few as 10 to 20 clusters. Other studies have advocated for Bayesian methods when the number of clusters is small (e.g., Baldwin & Fellingham, 2013; Browne & Draper, 2006; Hox et al., 2012; Stegmüller, 2013; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015), particularly for estimates of the variance components (which likelihood methods often have difficulty estimating with few clusters; Ferron et al., 2009), and Gelman (2006) showed that the intercept variance can be estimated without bias with as few as three clusters if the prior distribution is carefully considered.

Despite the increasing amount of research that has been carried out in this area, two aspects have still been

CONTACT Daniel McNeish  d.m.n.mcneish@uu.nl  P.O. Box 80140, Utrecht, 3508 TC, the Netherlands.

 Supplemental data for this article can be accessed on the [publisher's website](#).

¹Referring to these methods collectively as “design-based methods” is not a technical classification and is used to colloquially distinguish between models that account for clustering with and without random effects. These methods may also be generally referred to collectively as Taylor series linearization or as population-averaged models.

© 2016 Daniel McNeish, and Laura M. Stapleton. Published with license by Taylor & Francis.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

largely unaddressed. First, the small-sample performance of these methods has not been simultaneously investigated and compared because there are very strong disciplinary preferences regarding the method by which clustering is accounted for. For example, Bauer and Sterba (2011) noted that 94% of psychology studies account for clustering with MLMs whereas Peterson (2009) found that only 3% of studies in economics account for clustering with MLMs. As a result, few studies compare methods from different classes of methods (e.g., MLMs vs. GEE). Studies have compared smaller, select subsets of methods such as the Kenward-Roger correction with MLMs with the Morel-Bokossa-Neerchal correction with GEE (McNeish & Harring, 2015); the Kauermann-Carroll, Mancl-DeRouen, and Morel-Bokossa-Neerchal corrections (Lu et al., 2007); Bayesian Markov chain Monte Carlo (MCMC) with Kenward-Roger (Baldwin & Fellingham, 2013); and MCMC with maximum likelihood (ML) and restricted ML for MLMs (Browne & Draper, 2006). However, studies have yet to concurrently compare MLMs, GEE, Bayesian methods, and fixed effect models (FEMs) that are common in econometrics for accommodating clustered data (e.g., Allison, 2005; Peterson, 2009). In addition, many previous studies have investigated statistical properties of the estimates such as coefficient bias and confidence interval coverage (which largely assesses the appropriateness of standard error estimates), but far fewer have examined the power across methods. Even though power will be limited due to the limited number of clusters, substantive researchers could benefit from knowing which types of methods yield greater relative power to maximize their chances of detecting non-null effects.

Second, few studies (see Ferron et al., 2009, or van de Schoot et al., 2015, for exceptions) are informative for what researchers should do in the common scenario of possessing data that have very few clusters (less than about 15). This analytic scenario is especially common in research conducted within schools (e.g., education, developmental psychology) because it can be quite difficult and expensive to recruit even 10 schools or classrooms to participate in a study, and each cluster often has a fairly large number of units (e.g., roughly 30 students in primary school classrooms), which consume financial resources very quickly. To date, there is very little consensus on the best way to approach this type of analytic situation. As evidence, an exchange on the JISC Multilevel Listserv from August 2013 featured many prominent statisticians and research methodologists who were unable to reach an agreement or provide an illuminating citation for the most advantageous method to model data with only 8 clusters.

To outline the remainder of this article, we will describe a motivating applied example from a study funded

by a grant from the Institute for Educational Sciences that featured clustered data coming from only 12 classrooms. We proceed by briefly overviewing 12 different methods for handling clustered data that span multiple disciplines. A simulation study is then provided to systematically compare these competing methods with few clusters and continuous outcomes. Of particular interest will be the regression coefficient estimates (for all models) and variance component estimates (for the select models where this information is included). Recommendations are then provided and the results discussed.

Motivating example

The motivation behind this article arose from a cluster randomized trial in educational psychology that, despite having a moderate number of students, had a very small number of clusters. The data are from an Institute of Educational Sciences Development Grant² that investigated the efficacy of a Reading Buddies intervention to assess whether a researcher-designed treatment applied at the classroom level affected students' reading vocabulary compared to students in a control group who did not receive the treatment. The example data that we will expound upon are used to address only one of several research questions posed within this project and included 203 kindergarten students clustered within 12 classrooms in a semiurban, mid-Atlantic school district. The outcome measure was students' posttest vocabulary scores (as measured by the Peabody Picture Vocabulary Test Growth Scale Value, PPVT-GSV; $M = 121.56$, $SD = 21.07$), which were predicted by treatment group status, English language learner (ELL) status, PPVT-GSV pretest score, and relevant interactions thereof. Inference on the regression coefficients was the primary interest, so a variety of methods were available to model these data.³ These data are of particular interest because the inferential decision for multiple predictors is borderline significant as judged by a p value less than .05 or a Bayesian credible interval including 0, and the choice of method to model the data could greatly alter the interpretation of the effects, especially considering that the project was a development grant whose aim was to decide whether the intervention should be scaled up (requiring a significant time and financial investment). Furthermore, as alluded to previously, there are no extant studies that compare the various small-sample methods to one another in a way that would

² The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A110142 to the University of Maryland. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

³ If cluster-specific inferences or partitioning of the error variance in the individual- and classroom-level components had been the primary interest, MLMs would have been the only appropriate modeling choice for these data.

be informative to discern which method's estimates and resulting inferences are the most trustworthy.

Although some studies that were mentioned in the previous section include simulation conditions for the 12 clusters obtained in this data set, methods are scarcely compared directly to one another, and many data-generation models in these studies are rather simple and feature a single continuous variable at each level. It has been noted in previous small-sample research that model complexity increases sample size demands (e.g., McNeish & Stapleton, 2014). As a result for this particular analysis, despite the growing literature on accommodating small-sample clustered data, few studies were informative for the best practice to model these data. In addition, because the outcome was continuous and cluster-specific estimates are not desired, total effects as estimated by MLMs, GEE, or FEMs are equally interpretable, but no studies outside of a technical report by Schochet (2015) have attempted to systematically compare these varying frameworks in the context of few clusters.

We will first review the competing methods to account for clustering because some methods are essentially invisible in some literatures despite being very common in others. We will then present results from a simulation study to explore which methods minimize estimation bias, Type I error rate, and power concerns with few clusters. We conclude by analyzing these motivating data with the 12 competing methods to show how estimates compare across methods.

Overview of competing methods

Multilevel models

To account for clustering, MLMs directly model the clustering with random coefficients. Regression coefficients in an MLM consist of two possible types of effects: a fixed effect and a random effect. Fixed effects represent the relation between a predictor and the outcome regardless of the cluster affiliation of the observation, similar to coefficients in a standard single-level regression model. For each cluster, a cluster-specific random effect may be included (but is not required for all coefficients). A random effect captures how much the relation between the predictor and the outcome differs from the fixed effect estimate within a particular cluster.

Notationally, the model can be written as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j, \quad (1)$$

where \mathbf{y}_j is an $m_j \times 1$ vector of responses for cluster j ; m_j is the number of units within cluster j ; \mathbf{X}_j is an $m_j \times p$ design matrix for the predictors in cluster j (at

either level in this notation); p is the number of predictors (which includes the intercept); $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed regression coefficients; \mathbf{Z}_j is an $m_j \times q$ design matrix for the random effects of cluster j ; q is the number of random effects ($p \geq q$); \mathbf{u}_j is a $q \times 1$ vector of random effects for cluster j ; $E(\mathbf{u}_j) = \mathbf{0}$, and $Cov(\mathbf{u}_j) = \mathbf{G}$, where \mathbf{G} is $q \times q$, and $\boldsymbol{\varepsilon}_j$ is an $m_j \times 1$ vector of residuals of the observations in cluster j where $E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$, $Cov(\boldsymbol{\varepsilon}_j)$ is $m_j \times m_j$ and it is often assumed that $Cov(\boldsymbol{\varepsilon}_j) = \mathbf{R}_j = (\sigma^2\mathbf{I})$ for cross-sectionally clustered data, and \mathbf{u}_j and $\boldsymbol{\varepsilon}_j$ are independent ($Cov[\mathbf{u}_j, \boldsymbol{\varepsilon}_j] = \mathbf{0}$). The following subsections will discuss the basics of estimating these models with likelihood methods and Bayesian MCMC.

Likelihood estimation

The default estimation for MLMs with continuous outcomes in most software routines (SAS Proc Mixed, the lme4 R package, HLM 7) is restricted maximum likelihood (REML), which is known to exhibit better finite sample properties compared to traditional maximum likelihood, especially for estimates of the elements of the \mathbf{G} matrix (e.g., Browne & Draper, 2006; Cheung, 2013; McNeish & Stapleton, 2014). Rather than estimate all parameters simultaneously as in traditional maximum likelihood, the variance components and fixed effects are estimated in different phases. At a basic level, first the residuals from OLS are obtained (ignoring possible variance components), which by definition are independent of the fixed effects and have a mean of 0. Then maximum likelihood is applied to these OLS residuals to estimate the variance components. Once the variance components are estimated, these estimates are used in a generalized least squares estimator for the fixed effects. Estimation iterates between the variance components and the fixed effects until convergence is reached. More specifically, the log-likelihood function for the variance components housed in the \mathbf{G} and \mathbf{R} matrices can be written up to a constant as

$$l_j^{\text{REML}}(\mathbf{G}, \mathbf{R}_j) = -\frac{1}{2} \log |\mathbf{V}_j| - \frac{1}{2} \log |\mathbf{X}_j^T \mathbf{V}_j^{-1} \mathbf{X}_j| - \frac{1}{2} (\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{\text{GLS}})^T \mathbf{V}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{\text{GLS}}), \quad (2)$$

where \mathbf{V}_j is the model-based variance of the outcome for cluster j such that $\mathbf{V}_j = \text{Var}(\mathbf{y}_j) = \mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^T + \mathbf{R}_j$, and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is the generalized least squares estimator of the fixed effects, $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$. The improved finite sample performance comes from the inclusion of the $\frac{1}{2} \log |\mathbf{X}_j^T \mathbf{V}_j^{-1} \mathbf{X}_j|$ term that accounts for the degrees of freedom lost in estimating $\boldsymbol{\beta}$, which is not included in the traditional log-likelihood formula, which

is formulated up to a constant for the j th cluster as

$$l_j^{ML}(\mathbf{G}, \mathbf{R}_j) = -\frac{1}{2} \log |\mathbf{V}_j| - \frac{1}{2} (\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{GLS})^T \mathbf{V}_j^{-1} \times (\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{GLS}). \quad (3)$$

Stata's mixed procedure, MLwiN, and *Mplus* use traditional maximum likelihood as the default estimation method although Stata's mixed procedure can implement REML via an optional command.

Asymptotically, $\boldsymbol{\beta}$ can be shown to be distributed $MVN(\hat{\boldsymbol{\beta}}, Var^{MLM}(\hat{\boldsymbol{\beta}}))$ where $Var^{MLM}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Phi}_{MLM} = -(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T})^{-1} = \{\sum_{j=1}^J (\mathbf{X}_j^T \hat{\mathbf{V}}_j^{-1} \mathbf{X}_j)\}^{-1}$ for l either l^{ML} or l^{REML} depending on the estimation scheme (Fitzmaurice, Laird, & Ware, 2004 p. 92; Raudenbush & Bryk, 2002, p. 59) because, by definition, the variance components are independent of the regression coefficients when normality is upheld (i.e., $E(\frac{\partial l^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\phi}}) = 0$, where $\boldsymbol{\phi} = Vec(\mathbf{G}, \mathbf{R})^T$; Jacqmin-Gadda, Sibillot, Proust, & Thiébaud, 2008).

Kenward-Roger correction

Although multiple small-sample corrections exist (e.g., Manor & Zucker, 2004; Skene & Kenward, 2010a, 2010b; Zucker, Liberman, & Manor, 2000), the Kenward-Roger (Kenward & Roger, 1997, 2009) is the most widely implemented and most accessible in mainstream software such as SAS or Stata (new in Stata 14 released in April 2015), and several studies have explored the properties of the Kenward-Roger correction (e.g., Bell et al., 2014; Ferron et al., 2009; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; McNeish & Stapleton, 2014; Spilke, Piepho, & Hu, 2005; Vallejo & Livacic-Rojas, 2005). The Kenward-Roger correction is rather complex mathematically, so we will conceptually describe it for the remainder of this section.

In general with a small number of clusters, there are two concerns with respect to the quality of model estimates: (1) $\hat{\boldsymbol{\Phi}}_{MLM}$ is susceptible to downward bias with a small number of clusters, and (2) the denominator degree of freedom approximations for inferential tests of regression coefficients can have a large effect on resultant p values. The effect of (1) is that standard errors will be too small, which will inflate the Type I error rate of inferential tests. Kenward and Roger (1997) noted that the small-sample bias is attributable to two sources: (a) $\hat{\boldsymbol{\Phi}}_{MLM}$ is a biased estimator with a small number of clusters, and (b) does not take into the account that there is variability in the estimates that are used to compute $\hat{\boldsymbol{\Phi}}_{MLM}$. The former had been addressed by Kackar and Harville (1984), who had used a Taylor series expansion around $\boldsymbol{\phi}$. Kenward and Roger (1997) incorporated and expanded upon Kackar and Harville's approximation, also through

Taylor series expansions. Thus, the first step in the Kenward-Roger correction is to eliminate bias from $\hat{\boldsymbol{\Phi}}_{MLM}$.

With (2), denominator degrees of freedom for regression coefficients in MLMs are often a contentious issue because the denominator degrees of freedom can only be exactly calculated under a handful of situations (i.e., completely balanced data with simple structures for \mathbf{G} and \mathbf{R} ; Schaalje, McBride, & Fellingham, 2002). For example, in SAS Proc Mixed or Stata mixed, users have the option of approximating degrees of freedom with five different methods, none of which are appropriate across all scenarios. With a large number of clusters, this issue is not necessarily vital because univariate inferential tests are asymptotically χ_1^2 distributed. However, with few clusters where F or t tests are used, even small differences in the denominator degrees of freedom can have a noticeable effect on p values. Thus, the second step of the Kenward-Roger correction provides a better denominator degree of freedom approximation through an augmented Satterthwaite-type procedure.

MCMC estimation

MCMC estimation has generally been considered advantageous with smaller samples because it does not rely on asymptotic sample sizes to produce unbiased estimates; it does not give inadmissible estimates (e.g., negative variances); and it does not require adjustments or corrections to the likelihood for diminished sample sizes (e.g., Hox et al., 2012). Briefly, MCMC treats parameters as random variables rather than fixed quantities as in frequentist methods. As a result, parameters in an MCMC analysis have posterior distributions rather than a single point estimate as in frequentist analyses (although the posterior distribution is frequently summarized with a measure of central tendency to obtain the analog of a point estimate). This posterior distribution is the combination of the likelihood (the same as used in frequentist analyses) and a prior distribution that is user specified before the model is run. For further expository details on MCMC estimation, readers are referred to Kruschke, Aguinis, and Joo (2012), van de Schoot et al. (2014), or Zyphur, Oswald, and Rupp (2015).

Despite the potential advantages of MCMC, one must carefully consider prior distributions with small samples, particularly for the variance components, because prior distributions have an increased effect on posterior distributions when sample sizes are smaller. Typical choices for uninformative priors for variance components include a uniform prior with a fairly large range for the standard deviation (Gelman, Carlin, Stern, & Rubin, 2003) or an inverse gamma prior with small positive hyperparameters

on the variance (Daniels, 1999). However, Gelman (2006) showed that these choices can actually be more informative than intended when the data have few clusters. Gelman found that uniform priors tend to overestimate the variance components and inverse gamma priors tend to underestimate the variance components. Gelman (2006) suggested using a half-*t* or half-Cauchy distribution (a Cauchy distribution is equivalent to a *t* distribution with 1 degree of freedom)⁴ for the variance components with few clusters. Using an applied example, he showed desirable performance using a half-Cauchy distribution with only three clusters. To date, although analytical arguments for half-*t* and half-Cauchy have been made (e.g., Polson & Scott, 2012), the performance (both absolute and relative to other priors) of these recommendations has not been systematically assessed.

Generalized estimating equations

Rather than explicitly modeling the clustering mechanism as is done with MLMs, GEE essentially view the model as a single-level model and apply statistical corrections (typically based on the so-called sandwich estimator; Huber, 1967; White, 1980) to produce standard error estimates (and parameter estimates as well in some cases, such as with binary outcomes) that account for the fact that data were clustered (Liang & Zeger, 1986; Zeger & Liang, 1986). The advantage of GEE is that the specification of the random effects and their covariance structure does not have to be explicitly modeled, meaning that there are far fewer assumptions required compared to MLMs (Zeger, Liang, & Albert, 1988).

To explicate the mathematical details, GEE is an algorithmic method to estimate generalized linear models that potentially violate the normality and/or independence assumption. Briefly, generalized linear models relate $E(\mathbf{y}_j | \mathbf{X}_j) = \mu_j$ to a linear predictor $\mathbf{X}_j \boldsymbol{\beta}$ through a link function $g(\cdot)$ (McCullagh & Nelder, 1989; McCulloch & Searle, 2001). In behavioral sciences, common link functions are the identity function for normally distributed outcomes, $g(\mu_j) = \mu_j$; the logit link for binary outcomes, $g(\mu_j) = \log(\mu_j / (1 - \mu_j))$; or the log link for count outcomes, $g(\mu_j) = \log(\mu_j)$. The variance of \mathbf{y}_j is then specified as $Var(\mathbf{y}_j) = v(\mu_j)\varphi$, where φ is a possibly unknown scale parameter ($\varphi = 1$ for binary and Poisson responses), and $v(\mu_j)$ is a known variance function [$v(\mu_j)$

$= \mathbf{I}_{m_j \times m_j}$ for normally distributed outcomes, $\mu_j(1 - \mu_j)$ for binary outcomes, and μ_j for Poisson distributed outcomes].

Liang and Zeger (1986) defined generalized estimating equations for the regression coefficients $\hat{\boldsymbol{\beta}}$ such that $\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j = \mathbf{0}$ where $\mathbf{D}_j = \mathbf{X}_j^T \mathbf{A}_j = \frac{\partial \mu_j}{\partial \boldsymbol{\beta}}$; $\mathbf{V}_j = \hat{\varphi} \mathbf{A}_j^{1/2} \mathbf{K}_j(\alpha) \mathbf{A}_j^{1/2}$ for $\hat{\varphi}$ a scale parameter estimated by $\hat{\varphi} = \frac{1}{N-p} \sum_{j=1}^J \sum_{i=1}^{m_j} e_{ij}^2$; $\mathbf{S}_j = \mathbf{y}_j - \mu_j(\boldsymbol{\beta})$ for \mathbf{y}_j an $m_j \times 1$ vector of outcomes for the *j*th cluster and $\mu_j(\boldsymbol{\beta})$ based up the regression coefficients; $\mathbf{A}_j = \text{Diag}[Var(\mu_{j1}), \dots, Var(\mu_{jm_j})]$; and \mathbf{K}_j is an $m_j \times m_j$ working correlation matrix comprising unknown parameters α that estimate the correlation of observations within clusters rather than it being explicitly modeled. The structure of \mathbf{K}_j is specified by the researcher a priori, but its elements are updated algorithmically. For cross-sectionally clustered data, an exchangeable structure is typically suitable⁵ where $Corr(Y_{ij}, Y_{kj}) = \begin{cases} 1 & i=k \\ \alpha & i \neq k \end{cases}$, meaning that an arbitrary within-cluster observation has equal correlation with all other observations within the same cluster. The value of α with an exchangeable working structure is conceptually similar to the traditional intraclass correlation (ICC) as calculated with MLMs in an unconditional model (Wu, Crespi, & Wong, 2012).

GEE iteratively updates the parameters in the working structure, α . First, $\hat{\boldsymbol{\beta}}$ is estimated assuming independence. Then, $\mathbf{K}_j(\alpha)$ is estimated from the errors of the model that assume independence. The estimation of $\mathbf{K}_j(\alpha)$ depends on the working structure specified by the researcher. For an exchangeable structure that is typical with cross-sectional clustering (Horton & Lipsitz, 1999), $\hat{\alpha} = \frac{1}{\hat{\varphi}(N^*-p)} \sum_{j=1}^J \sum_{i < k} e_{ij} e_{ik}$ where $N^* = 0.5 \sum_{j=1}^J m_j(m_j - 1)$. Once a value(s) for $\hat{\alpha}$ is obtained, then \mathbf{V}_j can be calculated by $\mathbf{V}_j = \hat{\varphi} \mathbf{A}_j^{1/2} \mathbf{K}_j(\alpha) \mathbf{A}_j^{1/2}$; $\hat{\boldsymbol{\beta}}$ can then be updated once \mathbf{V}_j is estimated such that $\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + (\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1} (\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j)$ where *r* is the index for the iteration. When $r = 1$, $\hat{\boldsymbol{\beta}}_r$ houses the coefficient estimates under the independence assumption.

Once the iterative process has successfully converged, $Var^{GEE}(\hat{\boldsymbol{\beta}})$ is calculated using a sandwich estimator based upon the naïve estimator, $Var(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Phi}} = (\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1}$ (e.g., McCullagh & Nelder, 1989). The naïve estimator “sandwiches” a quantity that takes the clustering into account. In GEE, the middle term is formulated by

⁴ The probability density function of the *t* distribution is $\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} (1 + \frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$, where ν is the degrees of freedom and *x* is a random variable. When $\nu = 1$, the pdf reduces to $\frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}\Gamma(\frac{1}{2})} (1 + x^2)^{-\frac{1}{2}} = \frac{1}{\pi(1+x^2)}$, which is the probability density function of the standard Cauchy distribution.

⁵ Ballinger (2004) states that “(when) there is no logical ordering for observations within a cluster (such as when data are clustered within subject or within an organizational unit but not necessarily collected over time), an exchangeable correlation structure should be used” (p. 133).

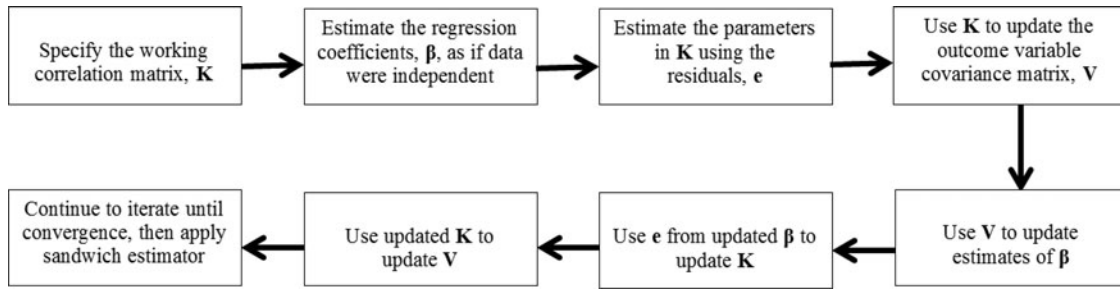


Figure 1. Conceptual flowchart of GEE algorithm.

$\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j$ making $\text{Var}^{GEE}(\hat{\boldsymbol{\beta}})$ equal to

$$\hat{\boldsymbol{\Phi}}_{GEE} = \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \times \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right) \times \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1}, \quad (4)$$

where the matrices have the following dimensions: \mathbf{D}_j is $m_j \times p$; \mathbf{V}_j is $m_j \times m_j$; and \mathbf{S}_j is $m_j \times 1$.

We realize that the GEE algorithm and the associated technical details may be rather opaque for some readers. To help assuage the technicality of this presentation, [Figure 1](#) shows a conceptual flowchart of the GEE algorithm. As a reminder, GEE avoids the complexity of modeling with random effects and, in essence, treats the random effects as a nuisance and does not attempt to estimate them. The resulting output of a GEE model looks nearly identical to a single-level model except that the GEE algorithm described earlier in this section adjusts estimates (coefficients, standard errors) for clustering. For a comprehensive treatment on differences between GEE and MLMs (the intricacies of which can be nuanced and are thus outside the scope of this article), readers are referred to either [Gardiner, Luo, and Roman \(2009\)](#) or [McNeish, Stapleton, and Silverman \(in press\)](#).

Small-sample GEE corrections

Similar to MLMs, the sandwich estimator for $\hat{\boldsymbol{\Phi}}_{GEE}$ that accounts for clustering in Equation (4) is consistent asymptotically; however, it is not unbiased when the number of clusters falls below about 40 (e.g., [Mancl & DeRouen, 2001](#); [Pan & Wall, 2002](#)). Two classes of small-sample corrected sandwich estimator have been proposed in the literature: residual-based corrections and design-based corrections. Residual-based corrections account for

small-sample bias by adding a matrix (or two depending on the correction) to the innermost part of middle term in the sandwich estimators (adjacent to the residual matrix, hence the term residual-based correction). These matrices inflate the standard error estimates, which are known to be downwardly biased with few clusters. Residual-based corrections rewrite the sandwich estimator from Equation (4) such that

$$\hat{\boldsymbol{\Phi}}_{RBC} = \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \times \left(\sum_{j=1}^J \omega_j \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{F}_j^T \mathbf{S}_j \mathbf{S}_j^T \mathbf{F}_j \mathbf{V}_j^{-1} \mathbf{D}_j \omega_j \right) \times \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1}. \quad (5)$$

Note that two matrices have been added in Equation (5) compared to Equation (4): \mathbf{F}_j and ω_j where ω_j is $p \times p$ and \mathbf{F}_j is $m_j \times m_j$. For the classic sandwich estimator, \mathbf{F}_j and ω_j are identity matrices and are thus not included in Equation (4). However, to correct for small-sample bias, various corrections have proposed different values for \mathbf{F}_j and ω_j .

[Table 1](#) summarizes the values of \mathbf{F}_j and ω_j used in these approximations.

The Morel-Bokossa-Neerchal correction ([Morel et al., 2003](#)) is the primary design-based small-sample correction employed in applied studies. Design-based

Table 1. Residual-based small-sample corrections to the sandwich estimator.

Correction	ω_j	\mathbf{F}_j
No Correction	\mathbf{I}	\mathbf{I}
Fay-Graubard	$\text{Diag}\{(1 - \min\{c, [\mathbf{Q}]_{jj}\})^{-1/2}\} \mathbf{I}$	\mathbf{I}
Kauermann-Carroll	\mathbf{I}	$(\mathbf{I} - \mathbf{H}_j^T)^{-1/2}$
Mancl-DeRouen	\mathbf{I}	$(\mathbf{I} - \mathbf{H}_j^T)^{-1}$

Note: $\mathbf{H}_j = \mathbf{D}_j^T \hat{\boldsymbol{\Phi}}_{GEE} \mathbf{D}_j \mathbf{V}_j^{-1}$; $\mathbf{Q} = \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \hat{\boldsymbol{\Phi}}_{GEE}$, $0 \leq c \leq 1$ where c is an upper bound for the correction, and diagonal values of ω_j cannot exceed 2. By default, SAS uses a value of $c = 3/4$.

corrections have the same desired result as residual-based corrections; however, design-based corrections take a different form and include additional additive terms to the classical sandwich estimator rather than appending matrices to the middle term. Specifically, the Morel-Bokossa-Neerchal correction is calculated by

$$\hat{\boldsymbol{\Phi}}_{MBN} = \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right) \times \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} + \delta \phi \left(\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1}, \quad (6)$$

where $\delta = \begin{cases} \frac{p}{(J-p)} & \text{if } J > (d+1)p \\ 1/d & \text{if } J \leq (d+1)p \end{cases}$ for p equal to the number of predictors in the model; J is equal to the number of clusters; d is a user-selected constant; and $\phi = \max(r, p^{-1} \text{tr}((\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1} (\sum_{j=1}^J \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j)))$. Common values for d and r that are also the SAS and geesmv R package defaults are 2 and 1, respectively.

Fixed effects model

With FEMs (a.k.a. dummy variable regression), cluster affiliation indicators (0/1 indicator variables, one for each cluster in the data) are included in the model as predictor variables with the goal being to account for the nested structure of the data without estimating the random effects, particularly when assumptions inherent with random effects are untenable or estimation may be computationally complex (Allison, 2005; Galbraith, Daniel, & Vissel, 2010). When indicators that represent cluster membership are added as predictors, the intercept is often removed from the model such that the cluster affiliation variables then represent the intercept value for each specific cluster, similar to how each cluster receives a cluster-specific intercept estimate in MLMs. Unlike MLMs, FEMs do not estimate random effects and thus require far fewer assumptions, which may be advantageous. With few clusters, FEMs also hold the added advantage that the cluster affiliation variables account for all heterogeneity at Level 2, allaying concerns about omitted variable bias at Level 2 that may occur if one has more potential predictors than degrees of freedom (as may occur with MLMs with very few clusters). Bias from omitted variables at Level 1 is still a concern, however.

Notationally, assuming the intercept term has been suppressed, the model can be written as

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \alpha_j C_j + \mathbf{r}_j, \quad (7)$$

where \mathbf{y}_j is an $m_j \times 1$ vector of responses for the j th cluster; \mathbf{X}_j is a $m_j \times p$ design matrix of substantive predictors

(there is no intercept); $\boldsymbol{\beta}$ is a $p \times 1$ vector of substantive regression coefficients; α_j is the cluster affiliation variable estimate for cluster C_j ; and \mathbf{r}_j is the residual that is traditionally assumed to be distributed $MVN(\mathbf{0}, \sigma^2 \mathbf{I})$.

A limitation of FEMs is that effects of Level 2 predictors⁶ cannot be estimated directly in the model although inclusion of Level 1 predictors or interactions between Level 2 and Level 1 predictors do not pose any problems in estimation (Allison, 2005; Gardiner et al., 2009; Murnane & Willet, 2010). Level 2 predictors and the cluster affiliation predictors will be perfectly collinear, meaning that both cannot be estimated simultaneously (Murnane & Willet, 2010). Instead, the effects of both measured and unmeasured variables at the cluster level are accounted for within the individual cluster affiliation coefficients (Allison, 2005; Murnane & Willet, 2010). This does present problems if a substantively relevant predictor is included at Level 2 (a common example would be a treatment effect in a cluster randomized trial as presented in the motivating example) because it too will be absorbed into the cluster affiliation coefficient estimates. However, if one assumes homogeneous slopes of Level 1 predictors across clusters (the equivalent of only a random effect for the intercept in an MLM), the treatment effect can be recovered using linear contrasts of the cluster affiliation variable coefficients. That is, one can inferentially test the treatment effect by taking a weighted average of the cluster affiliation estimates for the treatment group and comparing it to a weighted average of the cluster affiliation variable coefficient estimates for the control group. Mathematically, this can be expressed by calculating $\mathbf{L}\boldsymbol{\beta}$ where \mathbf{L} is a $1 \times p$ vector designating which effects to include, and $\boldsymbol{\beta}$ are the least squares coefficient estimates calculated by $(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$, whose naïve standard error is calculated by

$$\sqrt{\mathbf{L}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{L}^T \sigma^2}, \quad (8)$$

where σ^2 is the residual variance for the conditional model. However, as discussed in the next section, this naïve estimator will be inappropriately small when data are clustered.

Standard errors for Level 2 predictors

Although the effects for binary Level 2 predictors can be estimated through linear combinations of the cluster affiliation estimates under certain assumptions, the standard error estimates from Equation (8) will be too small

⁶ As a clarification, with FEM there is only a single level in the model, so the predictor does not enter the model at Level 2 as is conventional in MLMs. Rather, our terminology here indicates that the variable was collected at the second level. It may be helpful to conceptually label Level 2 predictors in FEMs as “cluster-level predictors.”

according to software computations (e.g., an ESTIMATE statement in SAS). As shown in Equation (8), the standard error estimates of a linear combination of coefficients is a function of the variance at the lowest level, σ^2 . In FEMs, σ^2 is not the total variance because the cluster affiliation dummies have accounted for the variance at Level 2. That is, whereas an MLM will consider variation at both Level 1 and Level 2 when calculating standard errors (i.e., $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$), FEMs do not model the Level 2 variance and therefore have no such mechanism to partition the variance. Software, however, will not recognize that the variation attributable to the cluster affiliation dummy variables is akin to Level 2 variance in MLMs. The Level 2 variance in MLMs is unexplained variance, meaning that it contributes to the standard error calculations, but this variation is considered explained variance in an FEM. Therefore, σ^2 is analogous to Level 1 variance in MLMs, which will necessarily make standard error estimates too small because the calculation is only based upon a fraction of the appropriate variance. No recommendations could be found in the literature to rectify this issue.

In an attempt to remedy this issue, we propose that standard error estimates output by software for effects estimated with linear combinations of regression coefficients be multiplied by the DEFT, which is the square root of the unconditional design effect. In survey statistics, DEFT is a quantity that measures the degree to which the standard error of the mean is inflated due to the use of cluster sampling as compared to that for data from a simple random sample. For example, a DEFT of 2 means that the standard error will be twice as large in a model that accounts for clustering than in a comparable model that ignores clustering. The DEFT is calculated as

$$\text{DEFT} = \sqrt{1 + (m - 1) \times \text{ICC}}, \quad (9)$$

where m is the average cluster size and ICC is the intraclass correlation calculated from the unconditional model.⁷ If the ICC is 0 (i.e., data are not meaningfully clustered), then DEFT = 1 and the Level 1 residual variance is equal to the total residual variance. To correct the

standard error estimates for quantities not explicitly output by the model (e.g., a Level 2 treatment effect), the standard error estimates output by the software program (which only account for Level 1 variance) will be multiplied by the DEFT to account for the residual variance present at Level 2 that is accounted for by the cluster affiliation variables.

Readers familiar with DEFT may note that it is a univariate measure rather than a global measure for the model (like the ICC). The cluster affiliation predictors are the fixed effect equivalent of random intercepts in an MLM—each cluster receives its own unique estimate. Under the assumption of homogeneous slopes, all of the Level 2 variation is contained within the cluster affiliation predictors, which are essentially cluster-specific intercept (fixed effect) estimates. Thus, the unconditional DEFT under homogeneity of slopes is capturing how much the intercept standard errors would increase with cluster sampling versus simple random sampling. Because the cluster-affiliation variables (which can be thought of as cluster-specific intercepts) are the only coefficients included in the linear combination for calculating the treatment effect, the degree of underestimation can be directly quantified by the unconditional DEFT. Should homogeneity of slopes not be a tenable assumption, the DEFT correction method will fail.

Differences between methods to accommodate clustered data

Although MLMs, GEE, and FEMs are all able to yield estimates that allow for appropriate and trustworthy inferences to be made with nonindependent data, there are some research questions and research scenarios in which one model may or may not give pertinent information.

Specifically, if researchers are interested in cluster-specific information, then MLMs are the only modeling framework that is appropriate. Examples of cluster-specific information include partitioning the variance between levels, prediction or inference for specific clusters in the data, or examining contextual effects for specific clusters. Cluster-specific questions can similarly be addressed with FEMs; however, the inferences are only appropriate to the clusters in the data because clusters are specified as fixed effects. In MLMs, clusters are assumed to be a random sample of the broader population of clusters,⁸ and thus inferences are generalizable to the

⁷ An anonymous reviewer raised a valid point that many computational formulas for the ICC use variance components from MLMs that make the assumptions that clusters are randomly sampled, an assumption not present in FEMs. In this article, FEMs are largely presented as a competing method to MLMs for data with few clusters, so this issue may not be overly problematic because the assumption would be met if MLMs were considered from the start. In addition, prior to widely accessible software for modeling clustered data, multiplying single-level, fixed effect model (without cluster-affiliation dummies) standard errors by the DEFT was a commonly recommended method to approximately account for clustering (e.g., Hahs-Vaughn, 2005; Huang, 2016; Thomas & Heck, 2001; Thomas, Heck, & Bauer, 2005) and the DEFT continues to be routinely used in inference from complex survey data (e.g., Lohr, 2014). None of these prior studies have noted any issues with the additional assumption of the ICC that clusters are randomly selected despite the fact that the FEM does not make this assumption.

⁸ The assumption that clusters are randomly sampled can be especially important when the data have few clusters because processes that are intended to be random may not be with small samples. Otherwise, the broad generalization of the results to clusters not included in the data may not be warranted (similar to FEMs).

Table 2. Summary of different information reported by MLMs, GEE, and FEMs.

	MLM	GEE	FEM
Covariance accounted by	Fully modeled with random effects	Working structure and cluster-robust estimator	Cluster affiliation dummy variables
SE Calculation	Information	Cluster-robust sandwich estimator	Closed form with OLS
Cluster-Specific inference	Yes and is generalizable to population	No	Yes but is restricted to clusters in the data
Partitions variance between levels	Yes	No	No
Number of clusters	Problematic with < 30 if uncorrected	Problematic with < 50 if uncorrected	Not consistent asymptotically

Note. MLM = multilevel model; GEE = generalized estimating equations; FEM = fixed effect model; OLS = ordinary least squares.

broader population rather than the finite sample of clusters as in FEMs. Consuming degrees of freedom is also an omnipresent concern with FEMs, and some of the aforementioned scenarios may require several additional parameters to be included in the model, which would make FEMs far less efficient than MLMs. GEE is strictly a population-average method and cannot make any inferences about specific clusters or partition the variance between levels. Contextual effects can be modeled with GEE (Begg & Parides, 2003; Berkhof & Kampen, 2004); however, the interpretation can only be made marginally. Table 2 provides a summary of the differences between MLM, GEE, and FEM.

These differences are quite meaningful with discrete outcomes and result in different interpretations of coefficients. That is, the inclusion of random effects fundamentally changes the interpretation of the regression coefficients in MLMs from a population-averaged interpretation inherent with single-level methods (a *population-averaged interpretation* is defined as follows: for a one-unit change in X , Y is predicted to change by β units, holding all other predictors constant) to a subject-specific interpretation (for a one-unit change in X , Y is predicted to change by β units, holding all other predictors constant *and the random effect values* constant). However, when the outcome is continuous, the interpretation between MLMs, GEE, and FEMs is identical⁹ because the random effects that are uniquely implemented with MLMs can be integrated out of the likelihood, meaning that the likelihood function is averaging over the random effects distribution, which yields the familiar population-averaged coefficient interpretation. Therefore, with continuous outcomes, there is much more flexibility regarding how one chooses to account for clustering because the interpretation across methods is the same. However, readers should note that this flexibility does not extend to discrete outcomes.

⁹ Because FEMs incorporate all observed and unobserved variability at Level 2 into the model, Level 1 coefficients may be conditional on different information compared to an MLM or GEE model if relevant Level 2 predictors are not measured or not included in the model. If all relevant Level 2 predictors are included in the model, then the Level 1 coefficients between FEMs, MLMs, and GEE will be the same.

Simulation study

Simulation design

To evaluate the performance of methods for modeling clustered data with few clusters, our simulation featured four conditions for the number of clusters (4, 8, 10, 14) and two conditions for the number of units within each cluster, which was set to be unbalanced according to what is commonly seen in practice (between 7 and 14 units per cluster; between 17 and 34 units per cluster). Keeping with the motivating example, the data-generation model consists of a continuous outcome variable (e.g., posttest scores) as a function of a binary variable with 50:50 prevalence at Level 2 (reminiscent of a treatment group assigned at Level 2), a continuous variable at Level 1 (X_{1ij} , reminiscent of a pretest score), a binary Level 1 variable with 50:50 prevalence (reminiscent of biological sex), and a binary Level 1 variable with 25:75 prevalence (reminiscent of English language learner status). In Raudenbush and Bryk (2002) notation,¹⁰ the generation model with hypothetical predictor variables can be formulated as

$$\begin{aligned}
 \text{Posttest}_{ij} &= \beta_{0j} + \beta_{1j}\text{Pretest}_{ij} + \beta_{2j}\text{Sex}_{ij} \\
 &\quad + \beta_{3j}\text{ELL}_{ij} + r_{ij} \\
 \beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Treatment}_j + u_{0j} \\
 \beta_{1j} &= \gamma_{10} + \gamma_{11}\text{Treatment}_j \\
 \beta_{2j} &= \gamma_{20} + \gamma_{21}\text{Treatment}_j \\
 \beta_{3j} &= \gamma_{30} + \gamma_{31}\text{Treatment}_j \\
 u_{0j} &\sim N(0, g_{00}), \quad r_{ij} \sim N(0, \sigma^2), \quad (10)
 \end{aligned}$$

where j is an index for the cluster ($j = 1, 2, \dots, J$), and i is an index for an observation within a cluster ($i = 1, 2, \dots, m_j$). The data-generation model only included Level 2 variation through the intercept (u_{0j}) because data with so few clusters would be unlikely to be able to support models of much greater complexity, and we did not wish to generate data from a model that would not be realistic to

¹⁰ Although the matrix form was presented earlier to facilitate discussion and estimation of models, we switch to Raudenbush and Bryk notation hereafter because Raudenbush and Bryk notation is better suited for discussing specific models because it more easily allows for readers to see which effects are located at which levels.

Table 3. Cohen's *d* population effect sizes for predictors in the data-generation model.

Parameter	Hypothetical effect	<i>d</i>
γ_{00}	Intercept	0.00
γ_{01}	Treatment	0.40
γ_{10}	Pretest	0.80
γ_{11}	Pretest \times Treatment	0.05
γ_{20}	Sex	-0.10
γ_{21}	Sex \times Treatment	0.02
γ_{30}	ELL	-0.30
γ_{31}	ELL \times Treatment	-0.20

fit under the circumstances of interest or that may have been fraught with convergence issues even if properly specified. We attempted to make the number of predictors realistic in terms of the quantity and level placement, in contrast to previous small-sample studies, which typically include a single continuous predictor at each level. We set the variance of the intercept random effect (g_{00}) to 1.625 and the residual variance (σ^2) to 3.00 across all conditions, resulting of an ICC of 0.20 in accordance with common ICC values in educational psychology research (the area of application motivating the study) seen in practice (Hedges & Hedberg, 2007).¹¹ Population values for regression coefficients were based on eta squared and then converted to Cohen's *d* as outlined in Fritz, Morris, and Richler (2012) and are presented in Table 3 for ease of interpretation. We chose approximate values based on what the predictors were intended to represent and chose to represent a range of different effect sizes.

The generated data were then fit with the 12 possible methods reviewed previously; Table 4 lists these methods, hyperparameters for prior distributions (if applicable), associated SAS procedures, and additional software frequently used by behavioral scientists that is capable of implementing each method (note that for more basic methods such as an MLM with ML or MCMC with an inverse gamma prior, the software listed may not be exhaustive). All data were generated with Proc IML in SAS 9.3 and subsequently analyzed with Proc Mixed, Proc MCMC (which uses Metropolis-Hastings), Proc GLM, and Proc Glimmix. Annotated SAS code for running each of these 12 models is provided in Appendix B. Although Proc Genmod or the newly released Proc GEE is typically used to fit GEE models with quasi-likelihood methods in SAS, Proc Glimmix is the only

SAS procedure that contains the small sample corrections as preprogrammed procedures that are of interest in this study. Therefore, the covariance parameters in the GEE models are estimated with restricted maximum likelihood rather than the more common method of moments as outlined in Liang and Zeger (1986). For more detail on this difference, readers may consult Example 38.12 in the SAS 9.2 User's Guide (SAS Institute Inc., 2008).

Because stationarity is an important issue to consider with MCMC, we ran test replications using a different number of burn-in iterations, recorded iterations, and thinning to determine the optimal number to use across the simulation conditions. Using 10,000 burn-in iterations, 50,000 recorded iterations, and thinning by 50 was found to provide nonsignificant Geweke's tests for all parameters and autocorrelations with magnitude below 0.10 for all lags beyond lag 2. From findings in previous studies by Browne and Draper (2006) and Gelman (2006), the posterior distribution of the inverse gamma prior and half-Cauchy conditions will be summarized with the median, and the posterior distribution of the uniform distribution will be summarized by the mode. GEE used an exchangeable working structure, which is recommended when data are clustered cross-sectionally (Ballinger, 2004; Horton & Lipsitz, 1999). The exchangeable working structure should be a proper specification because, with continuous outcomes, GEE with an exchangeable working structure is equivalent (barring differences in estimation methods) to an MLM with random intercepts (Twisk, 2004).

Outcome measures

Three outcome measures were tracked and reported. First, the median relative bias was recorded for each parameter estimate to examine how well each method was able to estimate effects under such extreme sample sizes. Using criteria from Flora and Curran (2004), estimates with a magnitude of bias greater than 10% were considered meaningfully biased. Second, because a major concern with a small number of clusters is downwardly biased standard error estimates, which leads to inflated Type I error rates, we tracked the 95% confidence/credible interval coverage rate. From criteria in Bradley (1978), confidence/credible interval coverage rates between [0.925, 0.975] will be considered to be reasonably close to the nominal rate, suggesting adequate Type I error rates. Note that some of the small-sample corrections (e.g., Kenward-Roger) also adjust degrees of freedom, which will affect *t* statistics used in the computation of confidence intervals with frequentists methods. Last, empirical statistical power for each effect was documented, given that our aim

¹¹ Readers may note that using the traditional formula for the ICC, where $ICC = g_{00}/g_{00} + \sigma^2$, will not yield a value of 0.20 with the specified values. However, the ICC is based on an unconditional model such that the variance explained by the predictors is lumped in the error terms. After considering the variance explained by the predictors, the specified values for the variance components yield an ICC of 0.20.

Table 4. Twelve analysis models used in the simulation and associated software options.

Model	Estimation	Correction/prior	SAS Proc	Additional notable software
Multilevel model	ML	—	Mixed/Glimmix	Mplus, SPSS, Stata, MLwiN, R, HLM
	REML	—	Mixed/Glimmix	SPSS, Stata, MLwiN, R, HLM
	REML	Kenward-Roger	Mixed/Glimmix	Stata, R (pbkrtest)
	MCMC	$\Gamma^{-1}(0.01, 0.01)$	MCMC	Mplus, JAGS, STAN, WinBUGS, MLwiN
	MCMC	$U(0, 10)$	MCMC	Mplus, JAGS, STAN, WinBUGS, MLwiN
	MCMC	Half-Cauchy (0, 4)	MCMC	JAGS, STAN, WinBUGS
GEE	GEE	—	Genmod/Glimmix	SPSS, Stata, R (gee,geepack)
	GEE	Mancl-DeRouen	Glimmix	R (geesmv)
	GEE	Kauermann-Carroll	Glimmix	R (geesmv)
	GEE	Fay-Graubard	Glimmix	R (geesmv)
	GEE	Morel-Bokossa-Neerchal	Glimmix	R (geesmv)
Fixed effects model	OLS	—	GLM/Reg	Too numerous to list

Note: GEE = generalized estimating equation; MCMC = Markov chain Monte Carlo; ML = maximum likelihood; OLS = ordinary least squares; REML = restricted maximum likelihood. Generalized estimating equation models were fit with a compound symmetric working covariance structure. The hyperparameters of the uniform and half-Cauchy priors are rather small because they are applied to the standard deviation, not the variance.

is to make recommendations for which method(s) provide the greatest relative power under such extreme circumstances.¹²

Results

Parameter relative bias

Regression coefficient estimate bias

For the most part, there was very little bias observed in the estimates of regression coefficients across conditions. Frequentist MLMs and GEE underestimated the cross-level interaction and the treatment effect with four clusters. For all other parameters in all other conditions, the bias was less than ±10%. Full results for the 7–14 cluster-size condition are presented in Table 5. Because many of the methods under investigation in this study are corrections for appropriate inference, they do not affect the regression coefficient estimation. Thus, Table 5 only shows frequentist MLMs estimated by ML and REML, Bayesian MLMs, classical GEE, and FEMs. Results for the 17–34 cluster-size condition were similar (although slightly better) and are not reported to avoid redundancy.

Variance component estimate bias

Table 6 reports the variance component bias for the intercept random effect and the Level 1 residual. Only 6 of the 12 methods under investigation estimate Level 2 random effects, so estimates from FEMs and GEE are not reported in Table 6. In addition, REML and the Kenward-Roger correction produce the same variance component estimates, so they are merged into a single column. As can be expected from prior research (e.g., Browne & Draper,

2006), the ML intercept variance estimate was highly negatively biased for all conditions of the simulation. Furthermore, as discussed in Ferron et al. (2009) and McNeish and Stapleton (2014), REML vastly reduces the estimation bias in intercept variance. However, REML begins to falter at about 10 clusters once models become even moderately complex (Browne and Draper, 2006, found no discernable bias with as few as six clusters in a very simple model). With small samples and frequentist estimation, nonpositive definite covariance matrices are a

Table 5. Regression coefficient percent median bias by method for 10 or fewer clusters with 7 to 14 observations per cluster.

Clusters	Parameter	ML	REML	IG	Uni	HCchy	GEE	FEM
4	ELL	8	7	0	-1	-3	4	-6
	Pretest	1	1	0	0	-1	0	0
	Sex	-3	-5	9	3	3	-3	-7
	Sex × Treat	-5	-5	-8	-9	-6	-5	-7
	Treat	-15	-14	3	-2	7	-14	-5
	ELL × Treat	-55	-53	7	1	4	-36	-1
	Pre × Treat	-12	-12	1	4	7	-10	8
	Intercept	0	0	0	0	0	0	0
8	ELL	1	0	-1	0	-4	2	-3
	Pretest	1	1	1	1	0	1	0
	Sex	9	9	8	14	12	8	6
	Sex × Treat	-5	-5	-5	-6	-2	-5	-7
	Treat	-2	-2	0	0	1	-2	-3
	ELL × Treat	-6	-6	-3	-3	-1	-7	-1
	Pre × Treat	3	-3	-2	-2	0	-3	-4
	Intercept	0	0	0	0	0	0	0
10	ELL	0	-1	0	1	-3	-1	6
	Pretest	0	0	0	1	0	0	-1
	Sex	5	5	8	16	12	15	-1
	Sex × Treat	-3	-3	-2	-3	0	-3	-4
	Treat	0	0	1	1	0	0	2
	ELL × Treat	0	0	-1	0	10	0	-4
	Pre × Treat	-2	-2	-2	-3	-1	-2	-3
	Intercept	0	0	0	0	0	0	0

Note: ML = maximum likelihood; REML = restricted maximum likelihood; KR = Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equation; FEM = fixed effect model. For the Intercept, Pretest × Treatment, and Sex × Treatment effects, absolute bias is reported instead of relative bias because the true effects were either zero or very close to zero. Bold entries indicate bias that exceeded the 10% threshold suggested by Flora and Curran (2004).

¹² Although not reported in text, Appendix A shows the comparison of efficiency for each method as well. These were excluded from the main text because efficiency is largely related to power, which is presented in the text.

Table 6. Percent relative bias of variance components.

Cluster size	Clusters	Parameter	ML	REML/KR	IG	UNI	HCchy
7 to 14	4	g_{00}	-85	-20	-50	52	58
			-55	-15	-40	11	-4
			-36	-11	-21	12	-9
			-26	-7	-12	12	-10
	8	σ^2	-18	-3	5	2	1
			-10	-1	3	2	2
			-6	0	4	3	2
			-5	0	3	3	3
17 to 34	4	g_{00}	-74	-31	-33	47	50
			-45	-13	-12	7	2
			-32	-9	-7	7	-7
			-24	-9	-5	9	-10
	8	σ^2	-7	-1	2	1	1
			-3	0	1	1	1
			-3	-1	1	1	1
			-2	0	1	2	2

Note: ML = maximum likelihood; REML/KR = restricted maximum likelihood/Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = half-Cauchy MCMC prior. ML, REML, and KR do not include nonconvergent replications. GEE and FEM are not shown because they do not estimate variance components. In accordance with Browne and Draper (2006), the posterior with an inverse gamma prior was summarized by the median, and the uniform prior was summarized by the mode. Congruent with Gelman (2006), the posterior with a half-Cauchy prior is summarized by the median. Bold entries indicate bias that exceeded the 10% threshold suggested by Flora and Curran (2004).

common concern. Table 7 shows the percentage of non-definite covariance matrices across conditions. These replications are excluded from the results reported for the remainder of the article.

As expected according to Gelman (2006), MCMC with a uniform prior in this simulation resulted in very highly upwardly biased intercept variance estimates that became less biased as the number of clusters and the cluster size increased (although the choice of hyperparameters would of course influence these results to some degree). Unexpected according to findings in Gelman (2006) and Polson and Scott (2012), although using a half-Cauchy prior resulted in more desirable performance as compared to using a uniform prior, the bias in the intercept variance was still rather high for the smallest number of cluster conditions included in this study and was more or less

Table 7. Percentage of nondefinite covariance matrices by condition.

Number of clusters	Cluster size	ML	REML/KR
4	7 to 14	42	24
	17 to 34	25	14
8	7 to 14	18	9
	17 to 34	4	3
10	7 to 14	8	3
	17 to 34	1	0
14	7 to 14	2	1
	17 to 34	0	0

Note: ML = maximum likelihood; REML/KR = restricted maximum likelihood/Kenward Roger. The Kenward-Roger correction affects fixed effect standard errors and denominator degrees of freedom, so the variance component estimates are identical to standard REML estimation.

on par with an inverse gamma prior. With smaller cluster sizes, however, the half-Cauchy prior performed best, with the Kenward-Roger correction not too far behind. With larger cluster sizes, the inverse gamma prior performed approximately equal to Kenward-Roger correction. Overall, the half-Cauchy prior produced the best estimates of the variance components with few clusters although it appears that performance with very few clusters is adversely affected when the model has several predictors (as opposed to the model used in Gelman, 2006).

Confidence/credible interval coverage

Table 8 shows the confidence/credible interval coverage rates for all regression coefficients in the model for all 12 methods for the 7 to 14 cluster-size condition. The confidence/credible interval coverage rates for the 17 to 34 cluster-size condition were rather similar to the 7 to 14 cluster-size condition and are not reported for brevity.

Generalized estimating equations

Immediately in Table 8, it can be seen that classical GEE, the Kauermann-Carroll correction, and the Fay-Grubard correction do not perform well, especially with 10 or fewer clusters, and are at risk for highly inflated Type I error rates. Mancl-DeRouen had coverage rates that were much closer to nominal rates but were still a little too short, particularly for predictors at Level 1 with fewer than 10 clusters. Morel-Bokossa-Neerchal performed the best of all the GEE methods although the coverage rates tended to consistently be on the high end of Bradley's range, as was similarly found in McNeish and Harring (2015).

Fixed effect models

FEMs provided very good coverage rates for predictors directly estimated by the model regardless of the number of clusters. Before correcting the standard error estimates for the Level 2 variance, coverage rates were quite poor.¹³ However, the issue was anticipated and, after multiplying by DEFT, coverage rates were quite good and showed no evidence of deviating from the nominal rate.

Multilevel models

As has been demonstrated in previous research (e.g., Browne & Draper, 2006; McNeish & Harring, 2015), ML and REML tended to have coverage intervals that are shorter than the nominal rate, especially for predictors involving a variable at Level 2. Use of a Kenward-Roger correction was largely able to address this limitation and provided coverage rates within the nominal

¹³ Prior to the DEFT correction, the coverage rate for the treatment effect with 7 to 14 observations per cluster was 83% and with 17 to 34 observations per cluster it was 71% across all number-of-cluster conditions.

Table 8. Confidence/credible interval coverage of model parameters for cluster size of 7 to 14.

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	91	93	93	94	95	95	51	78	62	89	93	95
	Pretest	93	96	96	97	97	97	52	77	62	90	94	94
	Sex	93	96	96	96	96	96	52	78	62	87	94	94
	Sex × Treat	81	83	83	95	96	96	50	66	61	82	82	94
	Treat	75	80	82	97	99	100	74	83	79	87	88	93
	ELL × Treat	78	80	80	95	96	95	45	64	57	81	80	96
	Pre × Treat	82	84	84	96	97	97	48	66	59	83	83	96
Intercept	96	97	91	97	99	99	65	92	69	94	97	93	
8	ELL	93	95	95	95	95	95	67	82	78	88	94	93
	Pretest	94	96	96	96	96	95	70	82	78	90	94	95
	Sex	94	95	96	96	95	95	72	82	80	89	94	96
	Sex × Treat	92	92	94	96	96	96	75	85	83	92	93	96
	Treat	88	92	93	95	96	96	84	92	89	95	96	93
	ELL × Treat	91	92	93	95	95	95	72	84	82	92	93	95
	Pre × Treat	94	95	95	96	96	96	72	86	82	93	94	94
Intercept	95	97	95	96	96	96	78	88	83	92	97	93	
10	ELL	94	95	95	95	95	95	79	86	85	90	95	94
	Pretest	95	96	96	96	96	95	79	85	85	91	96	95
	Sex	95	96	96	96	95	96	82	87	86	91	96	96
	Sex × Treat	96	96	96	97	96	96	84	91	89	94	97	95
	Treat	92	95	96	96	95	95	91	95	93	96	98	94
	ELL × Treat	94	95	95	95	96	96	83	90	89	94	96	96
	Pre × Treat	94	95	95	95	96	95	81	89	88	94	96	93
Intercept	96	97	96	96	94	95	90	93	92	95	98	93	
14	ELL	94	95	95	95	95	95	86	89	89	92	97	94
	Pretest	94	95	96	95	95	95	85	89	89	92	96	96
	Sex	95	96	95	96	96	96	87	90	91	94	97	95
	Sex × Treat	95	95	96	96	96	96	88	93	93	96	98	95
	Treat	92	95	95	95	93	93	91	95	94	96	98	94
	ELL × Treat	94	95	96	95	95	95	88	93	92	95	97	95
	Pre × Treat	94	96	95	95	95	94	86	91	90	93	97	95
Intercept	95	95	95	94	93	93	90	92	92	94	98	93	

Note: KR = Kenward-Roger; IG = inverse gamma; MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FG = Fay-Graubard; KC = Kauermann-Carroll; MD = Mancl-DeRouen; MBN = Morel-Bokossa-Neerchal; FEM = fixed effect model. Bold entries indicate coverage intervals beyond [.925, .975] from Bradley (1978).

range except in the four-cluster condition. Although Ferron et al. (2009) generally found that a Kenward-Roger correction was able to estimate standard errors appropriately even for extremely small numbers of clusters, the data-generation model in this study was much larger, and so the slight dip in performance was anticipated (e.g., McNeish & Stapleton, 2014).

MCMC methods generally performed quite well, even in the four-cluster condition where many other methods tended to exhibit coverage intervals that were too short. The choice of the prior distribution with MCMC, however, was not arbitrary, and some choices yielded more desirable coverage intervals than others. The inverse gamma prior had coverage rates close to the nominal rate across all conditions; however, the uniform prior and half-Cauchy prior coverage rates were too wide for the Level 2 binary predictor with four clusters, which adversely affected power as will be discussed in the next section.

Power

Tables 9 and 10 show the empirical power rates for all regression coefficients in the model for all 12 methods for the 7 to 14 cluster-size and the 17 to 34 cluster-size

conditions, respectively. Unlike the previous section, both cluster-size conditions are reported because the difference in power was noticeable between the 7 to 14 and 17 to 34 conditions. Cells that are greyed out indicate that the confidence/credible interval coverage rates were outside the range recommended in Bradley (1978), and rejection rates are likely to be inappropriate as a result. In the following sections, we will discuss power in a relative manner; this is not intended to imply that data with 7 or 10 clusters is optimal or even sufficient from a design perspective.

Generalized estimating equations

Power for GEE, the Fay-Graubard correction, and the Kauermann-Carroll correction is almost completely uninterpretable because coverage rates were so poor. For conditions where one might reasonably expect to detect effects (i.e., where Cohen's d is 0.20 or larger), both the Mancl-DeRouen and Morel-Bokossa-Neerchal corrections had moderately less power than MLMs and FEMs. Although the Morel-Bokossa-Neerchal correction was the only GEE method to generally yield appropriate coverage rates, it appears that the price paid is diminished power. McNeish and Harring (2015) similarly found

Table 9. Empirical power of model parameters for the unbalanced cluster condition with 7 to 14 observations per cluster.

Clusters	Parameter	ES	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	0.30	22	16	16	12	11	11	66	42	54	24	20	11
	Pretest	0.80	96	95	95	88	88	88	100	97	100	81	97	99
	Sex	0.10	8	6	5	5	5	4	50	27	42	14	9	7
	Sex×Treat	0.02	19	18	17	5	4	4	50	34	40	18	18	6
	Treat	0.40	43	33	26	10	2	2	40	26	31	19	18	39
	ELL×Treat	0.20	25	22	22	7	7	8	59	41	46	22	25	7
	Pre×Treat	0.05	19	16	16	4	3	3	54	36	43	17	17	5
8	ELL	0.30	21	19	19	18	17	19	48	37	39	24	18	20
	Pretest	0.80	100	100	100	99	99	99	100	99	99	94	100	100
	Sex	0.10	9	7	7	7	7	7	31	21	24	13	8	7
	Sex×Treat	0.02	7	6	6	5	4	3	26	16	17	8	7	5
	Treat	0.40	43	35	29	27	22	29	42	26	29	16	17	61
	ELL×Treat	0.20	16	14	14	12	11	13	33	21	23	11	13	11
	Pre×Treat	0.05	16	14	14	12	11	13	33	21	23	11	13	11
10	ELL	0.30	28	26	26	25	26	23	43	36	37	28	22	27
	Pretest	0.80	100	100	100	100	100	100	100	100	100	99	100	100
	Sex	0.10	8	8	8	8	8	8	23	17	17	12	6	7
	Sex×Treat	0.02	5	4	4	4	4	4	15	10	11	6	4	5
	Treat	0.40	52	45	40	41	39	38	48	34	38	27	24	73
	ELL×Treat	0.20	17	15	14	14	13	13	29	18	20	12	11	13
	Pre×Treat	0.05	7	7	7	6	6	7	20	11	13	6	5	8
14	ELL	0.30	36	34	34	34	34	31	44	39	39	33	26	35
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	100
	Sex	0.10	9	9	9	8	8	10	17	14	14	11	7	9
	Sex×Treat	0.02	5	4	4	4	4	5	12	7	8	4	2	6
	Treat	0.40	66	60	56	58	59	58	61	50	53	45	38	81
	ELL×Treat	0.20	18	17	17	17	16	16	28	19	21	15	12	17
	Pre×Treat	0.05	8	7	7	7	7	6	17	12	13	7	4	7

Note: ES = population Cohen's d effect size; ML = maximum likelihood; REML = restricted maximum likelihood; KR = Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FG = Fay-Graubard; KC = Kauermann-Carroll; MD = Mancl-DeRouen; MBN = Morel-Bokossa-Neerchal; FEM = fixed effect model. Grayed entries indicate coverage intervals beyond [.925, .975] from Bradley (1978) and therefore represent noncomparable/inappropriate power estimates.

disparate power between Kenward-Roger and Morel-Bokossa-Neerchal with few clusters.

Fixed effect models

Overall, power rates for FEMs were higher than for other methods, especially with very few clusters, while also being one of only two methods that was able to control the Type I error rate with as few as four clusters (the other being an MLM with an inverse gamma prior). This is related to improved efficiency of the OLS estimates, which is displayed in Appendix A. Briefly, the FEM model estimates exhibit less sampling variability than other methods with very few clusters, meaning that the standard errors are rightfully smaller (about 15%–20% with fewer than 10 clusters, see Table A2 for full detail). Efficiency for FEMs is comparable to other methods once the number of clusters reaches the mid-teens and power is essentially equivalent.

Multilevel models

In general, different types of MLMs performed fairly similarly with regard to power for those cells in which coverage rates were near the nominal level. Kenward-Roger and MCMC with an inverse gamma generally performed well and maintained appropriate coverage rates. As expected

from the wide coverage intervals, MCMC with a uniform prior had noticeably smaller power for the treatment effect across conditions, and the half-Cauchy prior had slightly smaller power than the inverse gamma prior when there were fewer than 10 clusters. Across conditions, with very few clusters, MLMs were consistently outperformed by FEMs in terms of detecting true non-null effects.

Analysis of motivating data

Returning to the motivating example, we modeled the IES Reading Buddies data with each of the 12 competing methods. These data featured 203 students clustered within 12 classrooms, meaning that each classroom had approximately 17 students (range = 12 to 24) and students were meaningfully nested within classrooms as evidenced by an ICC of 0.21 and an unconditional DEFT of 2.09. The continuous outcome variable, PPVT posttest score, is regressed on five predictors: Treatment Effect (at Level 2), ELL, PPVT pretest score, Treatment Effect × ELL, and Treatment Effect × PPVT pretest score. ELL and PPVT pretest score were grand-mean centered prior to being included in the model in accordance with recommendations in Enders and Tofghi (2007) because the primary interest was on the treatment effect located at Level 2. In

Table 10. Empirical power of model parameters for the unbalanced cluster condition with 17 to 34 observations per cluster.

Clusters	Parameter	ES	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	0.30	33	32	32	26	25	25	74	53	63	31	29	28
	Pretest	0.80	100	100	100	94	93	94	100	100	100	84	100	100
	Sex	0.10	9	8	8	8	7	8	53	35	46	16	11	7
	Sex×Treat	0.02	18	17	17	5	5	4	52	37	43	18	19	5
	Treat	0.40	55	42	29	15	4	6	43	28	34	16	18	65
	ELL×Treat	0.20	23	22	22	10	9	9	61	43	50	21	23	14
	Pre×Treat	0.05	20	19	19	6	6	6	54	38	45	19	20	5
8	ELL	0.30	42	41	41	40	40	39	65	55	57	42	38	42
	Pretest	0.80	100	100	100	99	99	99	100	100	100	96	100	1
	Sex	0.10	12	12	12	12	11	11	37	27	31	18	11	12
	Sex×Treat	0.02	7	6	6	5	5	5	22	15	17	7	7	6
	Treat	0.40	56	47	35	35	29	31	52	35	41	22	23	87
	ELL×Treat	0.20	22	20	20	19	18	19	45	30	35	19	17	22
	Pre×Treat	0.05	10	10	9	8	8	9	29	18	21	10	8	8
10	ELL	0.30	54	53	53	53	51	51	67	60	60	51	45	57
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	100
	Sex	0.10	16	15	15	15	14	14	31	25	26	19	12	13
	Sex×Treat	0.02	5	5	5	5	4	5	16	10	12	7	3	5
	Treat	0.40	67	62	54	53	55	58	64	51	55	43	38	91
	ELL×Treat	0.20	29	28	28	27	26	26	44	32	36	26	22	30
	Pre×Treat	0.05	9	9	9	9	9	8	18	12	13	10	7	10
14	ELL	0.30	69	68	68	68	67	67	76	70	70	63	57	69
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	100
	Sex	0.10	17	17	17	17	17	17	28	24	24	20	11	15
	Sex×Treat	0.02	5	5	5	5	5	5	11	8	8	6	4	5
	Treat	0.40	79	76	72	72	75	75	77	68	70	64	57	96
	ELL×Treat	0.20	40	39	39	39	36	37	50	41	44	36	30	41
	Pre×Treat	0.05	9	9	9	9	9	9	15	11	12	9	6	12

Note: ES = population Cohen's *d* effect size; ML = maximum likelihood; REML = restricted maximum likelihood; KR = Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FG = Fay-Graubard; KC = Kauermann-Carroll; MD = Mancl-DeRouen; MBN = Morel-Bokossa-Neerchal; FEM = fixed effect mode. Grayed entries indicate coverage intervals beyond [.925, .975] from Bradley (1978) and therefore represent noncomparable/inappropriate power estimates.

Raduenbush and Bryk notation, the MLMs could be formulated as

$$\begin{aligned}
 PPVT_{Posttest_{ij}} &= \beta_{0j} + \beta_{1j}(ELL_{ij}) + \beta_{2j}(PPVT_{Pretest_{ij}} - \overline{PPVT_{Pretest}}) + r_{ij} \\
 \beta_{0j} &= \gamma_{00} + \gamma_{01}Treatment_j + u_{0j} \\
 \beta_{1j} &= \gamma_{10} + \gamma_{11}Treatment_j \\
 \beta_{2j} &= \gamma_{20} + \gamma_{21}Treatment_j
 \end{aligned}
 \tag{11}$$

The GEE models can be similarly written by removing the random effect u_{0j} and substituting to yield a single-level equation. The FEM can also be similarly specified by removing the random effect, γ_{00} , γ_{01} (although this can be estimated with contrasts) and then adding cluster affiliation predictors. Because the scale of the outcome variable was larger than in the simulation, the priors will be changed slightly to maintain their intended uninformative nature. Specifically, the uniform prior will range from 0 to 500, and the scale of the half-Cauchy distribution will be 100 rather than 16. Similar to the simulation, the MCMC models use 10,000 burn-in iterations with 50,000 recorded iterations thinned by 50. The Geweke test was not significant for any parameter, and the autocorrelations were well behaved, meaning that there is reasonable evidence that MCMC chains reached convergence.

The resulting estimates are provided in Table 11. Because the FEM accounts for all observed and unobserved variables at Level 2, the FEM estimates are conditional on different information and are thus noticeably different from each of the other models. Most important, the treatment effect with the FEM was about half the other methods and was not statistically significant. This difference will be discussed further in the Discussion.

Of particular note is the wide amount of variation in the estimate of the intercept variance among the multi-level models (range: 5.00 to 9.56). In addition, the wide variation of statistical significance (or 0 not being in the credible interval for MCMC models) can be readily seen: MLMs identified four significant predictors at an alpha level of .05 whereas the Kauermann-Carroll correction, Mancl-DeRouen correction, and Morel-Bokossa-Neerchal correction (methods with less desirable performance in the simulation) only indicated two significant predictors. This particular data analysis has many effects that closely straddle a *p* value of .05 and is thus a good example of how choice of method with few clusters can markedly affect the interpretation if one adjudicates importance of predictors according to *p* values.

It should be noted that these are empirical data, and therefore population parameter values, or which model is

Table 11. Comparison of estimates and standard errors/posterior standard deviations from Reading Buddy data across all 12 methods.

Effect	Multilevel models						
	ML	REML	KR	IG	Uni	HCchy	
Intercept	126.10	126.00	126.00	126.10	125.90	126.10	
ELL	3.14 (2.33)	3.26 (2.38)	3.26 (2.43)	3.10 (2.38)	3.32 (2.50)	3.23 (2.35)	
Pretest	0.88** (0.06)	0.88** (0.06)	0.88** (0.06)	0.88 [†] (0.05)	0.88 ^{††} (0.06)	0.88 ^{††} (0.05)	
Treat	6.98** (2.37)	6.96** (2.54)	6.96** (2.56)	6.96 ^{††} (2.54)	6.96 ^{††} (2.71)	7.13 ^{††} (2.70)	
ELL × Treat	−6.68* (3.19)	−6.70* (3.25)	6.70* (3.29)	−6.57 [†] (3.24)	−6.84 [†] (3.31)	−6.82 [†] (3.22)	
Pre × Treat	−0.19** (0.07)	−0.19* (0.08)	−0.19* (0.08)	−0.19 ^{††} (0.07)	−0.19 ^{††} (0.08)	−0.20 ^{††} (0.07)	
Intercept var	5.00	7.17	7.17	8.30	9.56	8.87	
Residual var	62.69	63.99	63.99	65.32	65.01	63.99	
Effect	GEE and fixed effect model						
	GEE	FG	KC	MD	MBN	FEM	
Intercept	126.40	126.40	126.40	126.40	126.40	127.28	
ELL	2.48 (3.03)	2.48 (3.92)	2.48 (3.51)	2.48 (4.09)	2.48 (3.68)	3.97 (2.53)	
Pretest	0.87** (0.06)	0.87** (0.07)	0.87** (0.07)	0.87** (0.07)	0.87** (0.08)	0.88** (0.06)	
Treat	7.18* (2.48)	7.18* (3.06)	7.18* (2.79)	7.18* (3.17)	7.18* (3.06)	3.70 (2.82)	
ELL × Treat	−6.74* (3.38)	−6.74* (4.26)	−6.74 (3.88)	−6.74 (4.49)	−6.74 (4.38)	−6.91* (3.38)	
Pre × Treat	−0.17* (0.08)	−0.17 (0.10)	−0.17 (0.09)	−0.17 (0.10)	−0.17 (0.11)	−0.20** (0.08)	
Residual Var	69.72	69.72	69.72	69.72	69.72	63.99	

Note: Standard errors/posterior standard deviations appear in parentheses. To aid interpretation, the mean of the outcome was 123.56 with a standard deviation of 21.07.

[†] 95% credible interval does not contain 0.

^{††} 99% credible interval does not contain 0.

* $p < .05$; ** $p < .01$.

closest to “truth,” cannot be determined. In addition, due to space limitations, we do not report tests of the statistical assumptions inherent with each model (assumptions are overviewed in the Discussion).

Discussion

Very broadly, the 30-cluster / 30-unit recommendation for minimum sample size with clustered data that is often attributed to Kreft (1996) still permeates in much applied literature but is quickly being rendered obsolete, outdated, and inaccurate as methodological advances continue to burgeon. As shown in this study, for a moderately sized model, many methods are able to produce estimates with desirable properties with fewer than 10 clusters although the analysis will almost certainly be underpowered to some degree for any effects that are not large in magnitude. There are clear choices for which methods are preferable when one encounters data with few clusters and a moderate number of predictors, however.

First, estimating the model with uncorrected GEE is a poor choice as the standard error estimates are heavily downwardly biased. Furthermore, most small-sample corrections to the sandwich estimator in GEE were also rather ineffective under the conditions of this simulation with the exception of the Morel-Bokossa-Neerchal correction. However, the Morel-Bokossa-Neerchal correction tended to overcorrect in the conditions in this study, which was shown to adversely affect power (as has been shown previously in McNeish & Harring, 2015, with a more complex model). In substantive research contexts with few clusters, a loss of power is not a trivial matter

because power will already be diminished due to the small number of clusters.

Of the MLM methods investigated, MCMC estimation with an inverse gamma prior and MCMC estimation with a half-Cauchy prior were the best choices when broadly considering bias, power, and coverage intervals concurrently. The magnitude of the bias of MLMs with an inverse gamma prior and an MLM with a half-Cauchy prior was about equal although the inverse gamma prior tended toward being downwardly biased and a half-Cauchy prior tended toward being upwardly biased. In addition, the inverse gamma prior performed slightly better when the cluster size was smaller (7 to 14 observations per cluster) whereas the half-Cauchy prior performed slightly better with larger cluster sizes (17 to 34 observations per cluster). It should be noted that, in general, MLMs require a large number of assumptions and each of these assumptions were met by the data-generation process. With real data, the various assumptions of MLMs may not necessarily be upheld. In addition, with few clusters, the assumptions themselves are difficult to test and validate, so it can be unclear whether the assumptions are met. Furthermore, the ubiquitous Hausman specification test (Hausman, 1978) that is commonly used to assess the tenability of random effect model violations encounters problems with small sample sizes (Schreiber, 2008; Sheytanova, 2014).

Perhaps surprising to behavioral science researchers due to their scarce usage (outside of economics), FEMs performed extremely well for modeling data with few clusters and a moderate number of predictors. With very few clusters, the efficiency of the FEMs surpassed all other

methods, which helped to produce the maximal amount of power. Although Bayesian methods are often touted as being advantageous with smaller samples, FEMs vastly outperformed Bayesian methods in the simulation, especially in terms of power. For example, for the treatment effect with only four clusters and 17 to 34 observations per cluster, the empirical power for the half-Cauchy prior was 4% and the empirical power for the inverse gamma prior was 10%. Compare those values to the FEM whose empirical power was 32%. Although still far short of the 80% (arbitrary) cutoff applied in behavioral science, applied researchers would much rather have power near 30% than 4% or 10% provided that the regression coefficients are unbiased and Type I error rates are controlled (which was the case with FEMs in the simulation).

In FEMs, the regression coefficients were estimated without bias; the model makes a minimal number of assumptions; and concerns about omitted variable bias at Level 2 are alleviated. The last of these advantages can be particularly useful for research with few clusters. These studies often collect primary data (large-scale data sets would not likely feature so few clusters), and researchers may not always have the funds to collect several measures or may not have the insight a priori to note which variables at Level 2 should have been collected. In the motivating example, this was rather salient—11 of the methods identified the treatment as being significant; however, the FEM treatment effect was noticeably smaller and not statistically significant. As is common in small data sets, the number of measured variables was not highly extensive and MLMs and GEE are limited to the variables available in the data. FEMs can account for unmeasured Level 2 variables, and it seems plausible that an unmeasured Level 2 variable might have been related to the treatment effect in the motivating data, and after conditioning on this variable, the treatment effect was reduced.

The main drawback with FEMs is that Level 2 predictors cannot be explicitly included in the model because the cluster-affiliation variables account for all variation at Level 2. However, in cases where very few clusters are present, information at Level 2 is often not an explicit research interest. That is, when the number of clusters falls in the single digits, the research questions are often not overly concerned with specific effects at the cluster level, and the sample size would not likely be sufficient to make meaningful inferences about these effects. The motivating example on vocabulary demonstrated this common occurrence—the interest was on the performance of students and the students happened to be naturally clustered within classrooms. The classrooms and their characteristics did not play a large role in the broader research interests of the study—students were the primary interest and they happen to be naturally clustered within classrooms. This extends to other disciplines as well—in

medical and epidemiological studies the interest is very often on patients or individuals who happen to be clustered within hospitals or geographic areas. The characteristics of a hospital, for example, are important to take into account, but the magnitude of effects at the hospital level and/or their statistical significance may not always be directly relevant.

Although not the direct focus in this study, the adjustment made to the standard errors of the binary Level 2 predictor through the DEFT was quite effective, provided that some fairly rigid assumptions were upheld. None of the simulation conditions exhibited confidence interval coverage rates outside of the criteria in Bradley (1978). On the contrary, if the standard errors were uncorrected, the confidence interval coverage rates would have been in the high 60s to low 80s across conditions, which are clearly indicative of underestimated standard errors considering that the coefficient estimate was unbiased. Overall, this shows that FEMs may be of more utility than previously thought with very few clusters in cluster randomized trials though conventional wisdom precludes Level 2 predictors from the model.

As noted by an anonymous reviewer, data with few clusters could also be reasonably considered as a multiple-group structural equation model (MG-SEM). In this framework, a regression model would be specified, but the coefficient estimates and possibly the error variance would be freely estimated for each cluster (the group variable). If one were interested, the variance of coefficients could then be easily calculated across clusters, or more importantly, one could conduct significance tests to determine whether paths differ across groups. The variance of parameters calculated from an MG-SEM would likely be larger than the variance calculated from an MLM because MLMs will use empirical Bayes to shrink the random effect estimates for specific clusters (Hox, 2002). The FEM could be considered a special case of MG-SEM where the coefficients for Level 1 predictors and error variance are constrained to be equal across clusters but the intercept for each cluster is allowed to be different in each cluster.

As a notable limitation of the simulation, the values for the hyperparameters in the prior distributions of the MCMC conditions could have affected the results. The selected values were intended to be reasonably noninformative while also considering the scale of the outcome. That is, the priors featured a wide support over plausible values for the variance components without being naively noninformative by specifying a distribution such as an unbounded uniform prior, which can have deleterious effects on model estimates with small samples (McNeish, 2016). Despite this intention, with small samples, minor changes in the hyperparameters can affect posterior distributions. For example, the uniform prior for the

standard deviation of the random effects was bounded by $[0, 10]$ in the simulation, but one could argue that $[0, 5]$ may have been just as noninformative or that $[0, 10]$ was possibly too informative. Given the small sample sizes of interest in this article, changing these bounds would have affected the resulting posterior distribution. In accordance with sensitivity checks provided in Depaoli and van de Schoot (2016), we reran the model after adjusting the values for the hyperparameters in the prior distributions for the variance components in the applied example. The substantive conclusions for all parameters were not altered, and the relative change in the summaries of the posterior distributions (the Bayesian equivalent of point estimates) was at most a few percent which, according to criteria in Depaoli and van de Schoot (2016), would be classified as a small to moderate effect.

As extensions of this study, the present simulation study considered models with Level 2 variation induced through random intercepts. For models in which multiple random effects may be posited, multivariate prior distributions are likely necessary to ensure that the resulting MCMC draws produce a positive definite covariance matrix. The inverse Wishart distribution is a common prior distribution choice; however, this results in drawing values for variances from an inverse gamma distribution. Wand, Ormerod, Padoan, and Führwirth (2011) showed that one could create a half- t distribution from a mixture of inverse gammas, and it could be worthwhile to gauge whether the differences between inverse gamma and half-Cauchy generalize to the multivariate extension. In addition, given the strong performance of FEMs, it would be important to determine whether the treatment effect at Level 2 could still be estimated with linear combinations of the cluster affiliation coefficients. The unconditional DEFT-based standard error estimate correction will fail if slopes are heterogeneous, so a more clever and generalizable solution to the standard error problem would be valuable.

As a concluding remark based upon the overarching theme of this article, researchers may want to consider and draw from methods from other disciplines when faced with methodological challenges. Methodological work is published in a wide variety of outlets that may often include substantive journals with which behavioral science methodologists are not familiar. For the problem of interest in this article, extant methods common to the area of application performed decently but could be equaled or improved upon fairly readily by considering methods common to economics. Although there are many methodological problems in need of solutions in the behavioral sciences, sometimes a viable solution may already be available, albeit from a slightly different, non-behavioral science vantage point.

Article Information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: The collection of the motivating data was supported by Grant R305A110142 from the Institute of Educational Sciences, United States' Department of Education.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Gregory Hancock, Jeffrey Haring, Tracy Sweet, and Leigh Leslie for their comments on prior versions of this manuscript, Rebecca Silverman for permission to use her data, and the thoughtful comments of two anonymous reviewers and the editor which greatly improved the quality of the manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

- Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Raleigh, NC: SAS Institute.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- 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. doi:10.1037/a0030642
- Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7*, 127–150. doi:10.1177/1094428104263672
- Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16*, 373–390. doi:10.1037/a0025813
- Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine, 22*, 2591–2602. doi:10.1002/sim.1524
- Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you

- go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology*, 10, 1–11. doi:10.1002/sim.1524
- Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics*, 29, 201–218. doi:10.3102/10769986029002201
- Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, 31, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x
- Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1, 473–514. doi:10.1214/06-BA117
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29, 238–249. doi:10.1198/jbes.2010.07136
- Cheung, M. W. L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling*, 20, 157–167. doi:10.1080/10705511.2013.742404
- Daniels, M. J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27, 567–578. doi:10.2307/3316112
- Depaoli, S., & van de Schoot, R. (2016). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*. Advance online publication. doi:10.1037/met0000065
- Emrich, L. J., & Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, 41, 19–29. doi:10.1080/00949659208811388
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, 12, 121–138. doi:10.1037/1082-989X.12.2.121
- Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57, 1198–1206. doi:10.1111/j.0006-341X.2001.01198.x
- Ferron, J. M., Bell, B. A., Hess, M. R., 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
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004) *Applied longitudinal analysis*. Hoboken, NJ: Wiley.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491. doi:10.1037/1082-989X.9.4.466
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141, 2–18. doi:10.1037/a0024338
- Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The Journal of Neuroscience*, 30, 10601–10608. doi:10.1523/JNEUROSCI.0362-10.2010
- Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28, 221–239. doi:10.1002/sim.3478
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton, FL: CRC press. doi:10.1214/06-BA117A
- Gunsolley, J. C., Getchell, C., & Chinchilli, V. M. (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics-Simulation and Computation*, 24, 869–878. doi:10.1080/03610919508813280
- Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education*, 73, 221–248. doi:10.3200/JEXE.73.3.221-248
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1271.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87. doi:10.3102/0162373707299706
- Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician*, 53, 160–169. doi:10.2307/2685737
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, 6, 87–93. doi:10.18148/srm/2012.v6i2.5033
- Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*, 84, 175–196. doi:10.1080/00220973.2014.952397
- Huber, P. J. (1967, June). *The behavior of maximum likelihood estimates under nonstandard conditions*. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 1, pp. 221–233).
- Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., & Thiébaud, R. (2008). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, 51, 5142–5154. doi:10.1016/j.csda.2006.05.021
- Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853–862. doi:10.2307/2288715
- Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387–1396. doi:10.1198/016214501753382309
- 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 & Data Analysis*, 53, 2583–2595. doi:10.1016/j.csda.2008.12.013
- Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement*, 64, 224–242. doi:10.1177/0013164403260196
- Kreft, I. G. G. (1996). Are multilevel techniques necessary? *An overview, including simulation studies*. Unpublished manuscript, Los Angeles, CA: California State University

- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, 15, 722–752. doi:10.1177/1094428112457829
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22. doi:10.1093/biomet/73.1.13
- Lohr, S. L. (2014). Design effects for a regression slope in a cluster sample. *Journal of Survey Statistics and Methodology*, 2, 97–125. doi:10.1093/jssam/smu003
- Lu, B., Preisser, J. S., Qaish, B. F., Suchindran, C., Bangdiwala, S. I., & Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63, 935–941. doi:10.1111/j.1541-0420.2007.00764.x
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1, 86–92. doi:10.1027/1614-2241.1.3.85
- Maas, C., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58, 127–137. doi:10.1027/1614-2241.1.3.85
- Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, 46, 801–817. doi:10.1016/j.csda.2003.10.005
- Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small sample properties. *Biometrics*, 57, 126–134. doi:10.1111/j.0006-341X.2001.00126.x
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London, UK: Chapman and Hall.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.
- McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics*, 41, 27–56. doi: 10.3102/1076998615621299
- McNeish, D. M., & Harring, J. R. (2015). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics-Simulation and Computation*. Advance online publication. <http://dx.doi.org/10.1002/03610917.2014.983648>
- McNeish, D. M., & Stapleton, L. M. (2014). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*. Advance online publication. <http://dx.doi.org/10.1007/s10648-014-9287-x>.
- McNeish, D. M., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000078
- Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, 45, 395–409. doi:10.1002/bimj.200390021
- Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.
- Pan, W., & Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, 21, 1429–1441. doi:10.1002/sim.1142
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, 22, 435–480. doi:10.1093/rfs/hhn053
- Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7, 887–902. doi:10.1214/12-BA730
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 512–524. doi:10.1198/108571102726
- Schochet, P. Z. (2015). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs* (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Schreiber, S. (2008). The Hausman test statistic can be negative even asymptotically. *Journal of Economics and Statistics*, 228, 394–405. doi:10.1515/jbnst-2008-0407
- Sheytanova, T. (2014). Accuracy of the Hausman test in panel data: A Monte Carlo study. (Unpublished thesis) Örebro University, Sweden. <http://oru.diva-portal.org/smash/get/diva2:805823/FULLTEXT01.pdf>
- Skene, S. S., & Kenward, M. G. (2010a). The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator. *Statistics in Medicine*, 29, 2825–2837. doi:10.1002/sim.4073
- Skene, S. S., & Kenward, M. G. (2010b). The analysis of very small samples of repeated measurements II: A modified Box correction. *Statistics in Medicine*, 29, 2838–2856. doi:10.1002/sim.4072
- Spilke, J., Piepho, H. P., & Hu, X. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 374–389. doi:10.1198/108571105X58199
- Stegmuller, D. (2013). How many countries for multilevel modeling? A Comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, 57, 748–761. doi:10.1111/ajps.12001
- Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, 42, 517–540. doi:10.1023/A:1011098109834
- Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analyses. *New Directions for Institutional Research*, 127, 51–72.
- Twisk, J. W. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology*, 19, 769–776. doi:10.1023/B:EJEP.0000036572.00663.f2
- Vallejo, G., & Livacic-Rojas, P. (2005). Comparison of two procedures for analyzing small sets of repeated measures data. *Multivariate Behavioral Research*, 40, 179–205. doi:10.1207/s15327906mbr4002_2

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, 6, 1–13. doi:10.3402/ejpt.v6.25216

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: applications to developmental research. *Child Development*, 85, 842–860. doi:10.1111/cdev.12169

Wand, M. P., Ormerod, J. T., Padoan, S. A., & Fuhrwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6, 847–900. doi:10.1214/11-BA631

Westgate, P. M. (2013). A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. *Statistics in Medicine*, 32, 2850–2858. doi:10.1002/sim.5709

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838. doi:10.2307/1912934

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, 33, 869–880. doi:10.1016/j.cct.2012.05.004

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121–130. doi:10.2307/2531248

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44, 1049–1060. doi:10.2307/2531734

Zucker, D. M., Lieberman, O., & Manor, O. (2000). Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society: Series B*, 62, 827–838. doi:10.1111/1467-9868.00267

Zyphur, M. J., Oswald, F. L., & Rupp, D. E. (2015). Rendezvous overdue: Bayes analysis meets organizational research. *Journal of Management*, 41, 387–389. doi:10.1177/0149206314549252

Appendix A Efficiency of estimates across simulation conditions

Table A1. Standard deviation of regression coefficients for the 7 to 14 cluster size conditions.

Clusters	Effect	ML	REML	IG	Uni	HCchy	GEE	FEM
4	ELL	1.18	1.18	1.18	1.18	1.18	1.18	1.05
	Pretest	0.55	0.55	0.55	0.55	0.55	0.55	0.46
	Sex	1.01	1.01	1.01	1.01	1.01	1.01	0.9
	Sex×Treat	1.43	1.43	1.43	1.43	1.43	1.43	1.24
	Treat	1.69	1.69	1.69	1.69	1.69	1.69	1.6
	ELL×Treat	1.59	1.59	1.59	1.59	1.59	1.59	1.49
	Pre×Treat	0.72	0.72	0.72	0.72	0.72	0.72	0.62
	Intercept	1.25	1.25	1.21	1.21	1.2	1.25	1.12
	8	ELL	0.75	0.75	0.75	0.75	0.75	0.75
Pretest		0.33	0.33	0.33	0.33	0.33	0.33	0.31
Sex		0.66	0.66	0.66	0.66	0.66	0.66	0.58
Sex×Treat		0.95	0.95	0.95	0.95	0.95	0.95	0.84
Treat		1.16	1.16	1.14	1.15	1.15	1.16	1.12
ELL×Treat		1.08	1.08	1.08	1.08	1.08	1.08	1.01
Pre×Treat		0.49	0.49	0.49	0.49	0.49	0.49	0.42
Intercept		0.82	0.82	0.81	0.81	0.81	0.82	0.79
10		ELL	0.66	0.66	0.66	0.66	0.66	0.66
	Pretest	0.3	0.3	0.3	0.3	0.3	0.3	0.26
	Sex	0.56	0.56	0.56	0.56	0.56	0.56	0.52
	Sex×Treat	0.83	0.83	0.83	0.83	0.83	0.83	0.75
	Treat	1	1	1	1	1	1	1
	ELL×Treat	0.95	0.95	0.95	0.95	0.95	0.95	0.89
	Pre×Treat	0.43	0.43	0.43	0.43	0.43	0.43	0.38
	Intercept	0.69	0.69	0.69	0.69	0.69	0.69	0.69
	14	ELL	0.54	0.54	0.54	0.54	0.54	0.54
Pretest		0.24	0.24	0.24	0.24	0.24	0.24	0.22
Sex		0.46	0.46	0.46	0.46	0.46	0.46	0.44
Sex×Treat		0.65	0.65	0.65	0.65	0.65	0.65	0.62
Treat		0.85	0.85	0.85	0.85	0.85	0.85	0.83
ELL×Treat		0.76	0.76	0.76	0.76	0.76	0.76	0.73
Pre×Treat		0.34	0.34	0.34	0.34	0.34	0.34	0.33
Intercept		0.6	0.6	0.6	0.6	0.6	0.6	0.59

Note. ML = maximum likelihood; REML = restricted maximum likelihood; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FEM = fixed effect model.

Table A2. Standard deviation of regression coefficients for the 17 to 34 cluster size conditions.

Clusters	Effect	ML	REML	IG	Uni	HCchy	GEE	FEM
4	ELL	0.65	0.65	0.65	0.65	0.65	0.65	0.62
	Pretest	0.27	0.27	0.27	0.27	0.27	0.27	0.26
	Sex	0.52	0.52	0.52	0.52	0.52	0.52	0.54
	Sex × Treat	0.74	0.74	0.74	0.74	0.74	0.74	0.72
	Treat	1.48	1.48	1.45	1.46	1.46	1.48	1.41
	ELL × Treat	0.92	0.92	0.92	0.92	0.92	0.92	0.88
	Pre × Treat	0.37	0.37	0.42	0.42	0.42	0.37	0.37
	Intercept	1.05	1.05	0.97	0.95	0.96	1.05	1.01
8	ELL	0.47	0.47	0.47	0.47	0.47	0.47	0.4
	Pretest	0.19	0.19	0.19	0.19	0.19	0.19	0.17
	Sex	0.37	0.37	0.37	0.37	0.37	0.37	0.35
	Sex × Treat	0.55	0.55	0.55	0.55	0.55	0.55	0.49
	Treat	1.05	1.05	1.04	1.04	1.04	1.05	0.99
	ELL × Treat	0.66	0.66	0.66	0.66	0.66	0.66	0.59
	Pre × Treat	0.27	0.27	0.27	0.27	0.27	0.27	0.25
	Intercept	0.73	0.73	0.72	0.72	0.72	0.73	0.71
10	ELL	0.41	0.41	0.41	0.41	0.41	0.41	0.36
	Pretest	0.17	0.17	0.17	0.17	0.17	0.17	0.16
	Sex	0.33	0.33	0.33	0.33	0.33	0.33	0.32
	Sex × Treat	0.48	0.48	0.48	0.48	0.48	0.48	0.45
	Treat	0.91	0.91	0.91	0.91	0.91	0.91	0.88
	ELL × Treat	0.57	0.57	0.57	0.57	0.57	0.57	0.52
	Pre × Treat	0.23	0.23	0.23	0.23	0.23	0.23	0.23
	Intercept	0.64	0.64	0.64	0.64	0.64	0.64	0.63
14	ELL	0.33	0.33	0.33	0.33	0.33	0.33	0.31
	Pretest	0.14	0.14	0.14	0.14	0.14	0.14	0.13
	Sex	0.27	0.27	0.27	0.27	0.27	0.27	0.27
	Sex × Treat	0.4	0.4	0.4	0.4	0.4	0.4	0.39
	Treat	0.73	0.73	0.73	0.73	0.73	0.73	0.74
	ELL × Treat	0.47	0.47	0.47	0.47	0.47	0.47	0.43
	Pre × Treat	0.2	0.2	0.2	0.2	0.2	0.2	0.19
	Intercept	0.52	0.52	0.52	0.52	0.52	0.52	0.54

Note. ML = maximum likelihood; REML = restricted maximum likelihood; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FEM = fixed effect model.

Table A3. Standard error estimate percent median bias by method with 7 to 14 observations per cluster.

Clusters	Effect	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MBN	MD	FEM	FEM-DEFT
4	ELL	-21	-13	-12	-5	3	3	-65	-36	-48	-13	-16	-2	-2
	Pretest	-17	-9	-8	-4	4	4	-64	-33	-48	-10	16	-1	-1
	Sex	-14	-5	-4	0	0	0	-61	-28	-43	-4	-6	-5	-5
	Sex×Treat	-9	0	1	1	1	1	-59	-22	-41	-1	0	-4	-4
	Treat	-27	-8	-7	22	80	81	-55	-23	-39	-3	-15	-29	-4
	ELL×Treat	-9	0	1	-4	5	5	-62	-27	-42	-4	-13	0	0
	Pre×Treat	-8	1	2	1	7	7	-62	-26	-43	-1	40	2	2
	Intercept	-31	-13	-13	22	103	103	-56	-31	-41	-7	-13	-30	-4
8	ELL	-16	-11	-10	-4	-3	-3	-45	-24	-27	-6	-5	-5	-5
	Pretest	-12	-7	-6	-5	2	2	-44	-23	-26	-3	12	-2	-2
	Sex	-8	-3	-2	0	1	1	-37	-16	-19	4	8	3	3
	Sex×Treat	-7	-2	-2	3	2	2	-36	-9	-17	4	15	3	3
	Treat	-16	-4	-4	22	20	15	-33	-8	-16	5	7	-29	-6
	ELL×Treat	-10	-5	-4	-1	-1	-1	-40	-12	-19	0	5	-1	-1
	Pre×Treat	-11	-6	-5	0	2	2	-42	-17	-24	-2	28	-5	-5
	Intercept	-19	-7	-6	6	21	18	-37	-18	-22	3	-1	-28	-6
10	ELL	-10	-7	-6	-5	-1	-1	-29	-15	-14	4	5	-5	-5
	Pretest	-10	-6	-5	-5	3	3	-30	-17	-16	3	5	-1	-1
	Sex	-5	-1	-1	0	0	0	-25	-11	-11	10	6	2	2
	Sex×Treat	-2	2	3	4	3	3	-21	-2	-6	14	15	2	2
	Treat	-10	-1	0	6	6	6	-19	-1	-6	14	13	-29	-4
	ELL×Treat	-8	-5	-4	-1	-1	-1	-27	-6	-10	6	12	-2	-2
	Pre×Treat	-8	-5	-4	0	1	1	-30	-11	-14	4	16	-4	-4
	Intercept	-10	-1	-1	6	6	6	-20	-6	-7	15	9	-28	-7
14	ELL	-6	-4	-3	-3	0	0	-19	-8	-8	12	4	-1	-1
	Pretest	-9	-6	-6	-5	2	2	-23	-14	-13	7	0	0	0
	Sex	-1	2	2	3	-2	-2	-15	-5	-4	18	8	-1	-1
	Sex×Treat	-1	1	2	2	2	2	-14	0	-3	17	10	-1	-1
	Treat	-7	0	0	1	-3	-3	-13	0	-3	17	7	-29	-7
	ELL×Treat	-2	1	1	2	-1	-1	-14	1	-2	17	12	-3	-3
	Pre×Treat	-5	-3	-2	-2	0	0	-20	-6	-9	10	5	-3	-3
	Intercept	-10	-3	-3	-1	-4	-4	-16	-8	-7	14	2	-29	-6

Note: ML = maximum likelihood; REML = restricted maximum likelihood; KR = Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FG = Fay-Graubard; KC = Kauermann-Carroll; MD = Mancl-DeRouen; MBN = Morel-Bokossa-Neerchal; FEM = fixed effect model; FEM-DEFT = fixed effect model with DEFT correction. Bold entries indicate bias that exceeded the 10% threshold suggested by Flora and Curran (2004).

Table A4. Standard error estimate percent median bias by method with 17 to 34 observations per cluster.

Clusters	Effect	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MBN	MD	FEM	FEM-DEFT
4	ELL	-9	-6	-5	0	0	-1	-58	-35	-44	-3	4	-2	-2
	Pretest	-11	-8	-7	-6	-5	-1	-58	-37	-43	-3	36	-2	-2
	Sex	-13	-10	-9	-8	-8	3	-60	-39	-45	-8	6	2	2
	Sex×Treat	0	3	4	-2	-2	1	-55	-23	-38	5	39	-1	-1
	Treat	-31	-11	-11	27	62	79	-56	-27	-41	-5	2	-48	-6
	ELL×Treat	3	7	7	5	6	-2	-55	-22	-37	8	25	-3	-3
	Pre×Treat	-1	2	3	-3	-2	0	-56	-26	-39	5	78	-1	-1
8	Intercept	-36	-19	-19	29	97	113	-59	-40	-47	-13	-10	-46	-8
	ELL	-5	-3	-3	-1	-1	-3	-38	-21	-22	3	10	-3	-3
	Pretest	-7	-5	-4	-4	-3	-1	-39	-22	-23	2	17	-2	-2
	Sex	-7	-5	-5	-4	-5	-3	-37	-20	-21	3	9	-4	-4
	Sex×Treat	-1	1	1	1	2	-5	-31	-6	-14	9	25	-3	-3
	Treat	-17	-4	-4	13	22	15	-32	-9	-17	5	11	-46	-8
	ELL×Treat	-1	1	2	2	3	0	-34	-9	-17	7	20	2	2
10	Pre×Treat	-5	-3	-3	-3	-3	0	-36	-11	-19	4	28	-1	-1
	Intercept	-20	-8	-8	10	21	18	-37	-21	-23	1	3	-46	-7
	ELL	1	2	3	3	4	-1	-22	-9	-10	13	9	-1	-1
	Pretest	-4	-3	-3	-2	-2	-2	-23	-10	-10	10	11	-3	-3
	Sex	-8	-7	-6	-6	-6	-4	-26	-14	-15	5	2	-3	-3
	Sex×Treat	-2	0	0	0	1	-3	-20	-2	-7	12	14	2	2
	Treat	-10	-2	-1	7	9	4	-19	-2	-7	12	9	-47	-7
14	ELL×Treat	5	6	6	6	7	2	-17	3	-3	18	18	3	3
	Pre×Treat	-5	-4	-3	-3	-2	0	-22	-3	-8	9	16	-1	-1
	Intercept	-11	-2	-2	6	2	6	-21	-9	-10	11	5	-46	-8
	ELL	-3	-2	-1	-1	-2	-5	-17	-8	-8	13	2	-2	-2
	Pretest	1	2	2	2	2	-4	-12	-2	-2	19	9	-5	-5
	Sex	-4	-3	-3	-3	-3	-2	-16	-8	-8	13	2	0	0
	Sex×Treat	0	1	1	1	2	-2	-12	1	-2	18	10	3	3
14	Treat	-7	0	0	5	-5	-5	-11	1	-3	17	7	-46	-7
	ELL×Treat	1	1	2	2	2	1	-13	0	-3	17	8	-1	-1
	Pre×Treat	-2	-1	-1	-1	-1	0	-14	-1	-4	15	8	-4	-4
	Intercept	-9	-3	-2	3	-7	-3	-15	-6	-6	15	3	-47	-8

Note: ML = maximum likelihood; REML = restricted maximum likelihood; KR = Kenward Roger; IG = inverse gamma MCMC prior; Uni = MCMC uniform prior; HCchy = MCMC half-Cauchy prior; GEE = generalized estimating equations; FG = Fay-Graubard; KC = Kauermann-Carroll; MD = Mancl-DeRouen; MBN = Morel-Bokossa-Neerchal; FEM = fixed effect model; FEM-DEFT = fixed effect model with DEFT correction. Bold entries indicate bias that exceeded the 10% threshold suggested by Flora and Curran (2004).