

Biases in multilevel analyses caused by cluster-specific fixed-effects imputation

Matthias Speidel¹ · Jörg Drechsler¹ · Joseph W. Sakshaug^{1,2}

Published online: 24 August 2017

© The Author(s) 2017. This article is an open access publication

Abstract When datasets are affected by nonresponse, imputation of the missing values is a viable solution. However, most imputation routines implemented in commonly used statistical software packages do not accommodate multilevel models that are popular in education research and other settings involving clustering of units. A common strategy to take the hierarchical structure of the data into account is to include cluster-specific fixed effects in the imputation model. Still, this ad hoc approach has never been compared analytically to the congenial multilevel imputation in a random slopes setting. In this paper, we evaluate the impact of the cluster-specific fixed-effects imputation model on multilevel inference. We show analytically that the cluster-specific fixed-effects imputation strategy will generally bias inferences obtained from random coefficient models. The bias of random-effects variances and global fixed-effects confidence intervals depends on the cluster size, the relation of within- and between-cluster variance, and the missing data mechanism. We illustrate the negative implications of cluster-specific fixed-effects imputation using simulation studies and an application based on data from the National Educational Panel Study (NEPS) in Germany.

Keywords Hierarchical multiple imputation · Cluster-specific fixed-effects imputation approach · Multilevel imputation approach · Linear mixed model

✉ Matthias Speidel
matthias.speidel@iab.de

¹ Institute for Employment Research, Regensburger Strasse 104, 90478 Nuremberg, Germany

² The University of Manchester, Oxford Road, M13 9PL Manchester, UK

Introduction

Missing values are a common problem in survey data, which can lead to bias if the nonresponse is not properly taken into account by the analyst. A widely accepted strategy to deal with this problem is imputation, which is based on the idea that missing values are replaced with plausible values to produce a completed dataset on which standard analysis models can be applied by the analyst with no, or a less severe, nonresponse bias.

A procedure to take the uncertainty from imputation directly into account is *multiple imputation* (MI). With MI, values are not imputed just once, but $M \geq 2$ times. This leads to M datasets that need to be analyzed, each with the same method leading to M estimates of the parameters of interest and their standard errors. The final inference is obtained by using simple procedures to combine the different results (Rubin's combining rules, Rubin 1987). For applications of (multiple) imputation in educational research see, for example, the overview by Peugh and Enders (2004).

From a theoretical perspective, it is essential that the imputation model is congenial to the model used by the analyst to ensure unbiased results based on the imputed data. Broadly speaking, congeniality means that the model specifications of the imputation model and the analysis model are compatible, i.e., they should be based on the same modeling assumptions (see Meng 1994 and Kenward and Carpenter 2007 for more details). For example, if the analyst is interested in explaining the performance of students in a competence test and uses socio-economic status as one of the predictors, but this predictor is not used when imputing missing values in the competence test, the imputation model and the analysis model would be uncongenial. Therefore, an imputation method should always be developed keeping in

mind the assumed analysis model to be carried out on the imputed data.

These considerations also hold for hierarchical datasets. These are datasets in which individual measurements are grouped; for example, students observed within the same class or repeated measurements on the same individual. Such hierarchical datasets might be analyzed using multi-level models (see Goldstein 1987 or O’Connell and McCoach 2008 and the short review provided in the section “[Multi-level modeling](#)” of this paper). Thus, to ensure congeniality, multilevel models should also be used at the imputation stage. However, most of the statistical software packages that are commonly used for imputation such as SAS, SPSS, or Stata, do not provide imputation methods explicitly designed for hierarchical data. To our knowledge, the only tools that allow for multilevel imputation models are the external SAS macro `MMI_IMPUTE` developed by Mistler (2013), some multiple imputation routines in MPlus (Asparouhov and Muthén, 2010), the standalone software `REALCOM-IMPUTE` (Carpenter et al., 2011), which also offers interfaces for MLwiN and Stata, and the R packages `mice` (van Buuren et al., 2015), `pan` (Schafer, 2016), and `jomo` (Quartagno & Carpenter, 2016).

However, `mice` is limited to two levels of hierarchy and continuous dependent variables while all other imputation routines rely on the restrictive joint modeling approach. Joint modeling, which assumes a joint density for all variables with missing data, is especially problematic if the model of interest is a random slopes model, since unlike the sequential regression approach implemented in `mice`, the joint modeling approach cannot deal with missing data in the slope variables (Enders et al., 2016, see also Drechsler 2011 for a general discussion of the pros and cons of the joint modeling approach).

Due to the sparseness of suitable software, using cluster-specific fixed-effects imputation has been recommended in the literature (Diaz-Ordaz et al., 2016; Graham, 2009). This approach is carried out by including dummy variables, representing the cluster membership of the observations, into the data (see section “[Cluster-specific fixed-effects imputation](#)”). This imputation strategy is also endorsed on the FAQ website for the multiple imputation module in Stata (StataCorp, 2011). Since the cluster-specific fixed-effects approach is easy to implement using standard imputation software, it has been used for the imputation of missing values in hierarchical datasets (see for example, Brown et al., 2009; Clark et al., 2010; Zhou et al., 2016). Research about imputation in hierarchical settings has only been undertaken in recent years with the earliest papers on this topic focusing only on the impacts on global fixed-effects (the regression coefficients) inferences (Reiter et al., 2006; Taljaard et al., 2008; Andridge, 2011). In educational research, it is often the random effects themselves (or derivatives, such

as the intra-class correlation) that are of particular interest when measuring the school effect (Lenkeit, 2012; Nye et al., 2004; McCaffrey et al., 2004b). The impacts on random effects were addressed in later papers but the authors either only focused on random intercept models (Drechsler, 2015; Lüdtke et al., 2017; Zhou et al., 2016), or the evaluations were limited to running simulation studies without analytical derivations to identify which factors influence the bias observed in the simulation studies (van Buuren, 2011; Enders et al., 2016; Grund et al., 2016). In random intercept models, it is assumed that within a cluster, the average intercept deviates from the global intercept by a cluster-specific random value. For example, this could mean that in a class the students score on average four points higher on a math test than the average population of students. This is in contrast to a random coefficients model where the effect of a covariate, x on y , randomly deviates from the global effect; for example, if the performance x in a previous test has a higher effect in a class than on average.

To our knowledge, the impact on random effects if fixed-effects models with cluster-specific slopes are used for imputation has not yet been studied analytically, despite the demand for such research (Drechsler, 2015; Lüdtke et al., 2017; Grund et al., 2016). Our paper closes this research gap by comparing cluster-specific fixed-effects imputation and multilevel imputation and generalizing the evaluations to all types of random coefficient models. We derive analytically why the variance of the random effects in the analysis model is positively biased when a cluster-specific fixed-effects imputation model, instead of a multilevel imputation model, is used. Further, we find that beyond the three factors governing this bias that were already identified in Drechsler (2015) (for the special case of random intercept models), the bias also depends on the mean and variance of the observed data (which are governed by the missing data mechanism). We present support for these findings using simulation studies and a real data application.

The remainder of the article is organized as follows: Section “[Related research](#)” summarizes the findings from previous studies, highlights their limitations, and describes our contributions to fill these research gaps. Section “[Multilevel modeling](#)” summarizes the ideas behind multilevel modeling and introduces the relevant notation. The different imputation methods are described in section “[Imputation models](#)”. The following section compares the different imputation strategies analytically and derives which factors influence the bias in random effects-based inferences. The theoretical findings are confirmed using extensive simulations in the “[Simulation study](#)” section. In the “[Real data application](#)” section, we compare the results of the imputation methods on educational research data. Finally, in the “[Conclusion](#)” section we provide a summary of our findings with some practical guidance and provide an

outlook for further research on the topic of hierarchical data imputation.

Related research

As mentioned previously, research about imputation in hierarchical settings is relatively sparse. Reiter et al. (2006) illustrated that ignoring clusters in the imputation process can lead to biased analysis results for clustered sampling designs. They also illustrated that including cluster-specific fixed intercepts for each cluster in the imputation model will lead to conservative inferences for the global fixed effects in the analysis model, increasing the chances of type II errors. In substantive research, this could mean that some covariates are found to have no significant effect on the target variable, while in reality there is one, which would have been found if a proper imputation would have been conducted. Taljaard et al. (2008) compared several imputation routines in a cluster randomized trial setting (clustered randomized trials are typically analyzed using multilevel models but sometimes imputed based on a cluster-specific fixed-effects approach). They found that simple imputation routines (such as cluster mean imputation) can be a suitable choice, but is inferior in performance compared to a congenial (multilevel) imputation. Andridge (2011) also focused on cluster randomized trials. She showed analytically and empirically that the MI variance estimator for the global fixed effects will be conservative if cluster-specific fixed-effects imputation models are used. All three papers leave two kinds of research gaps. First, they limited their evaluations to random intercept models; and second, they all dealt with situations in which the random effects are only nuisance parameters. Thus, none of them evaluated the impacts of cluster-specific fixed-effects imputation on random effects inferences. However, as illustrated in the Introduction, these inferences are often of major interest in education research.

The first paper that also evaluated the impacts on random effects inferences is van Buuren (2011). In a simulation study, the author evaluated the consequences of ignoring the hierarchical structure completely or incorporating dummy variables for the clusters in a random intercept model. He found that ignoring the hierarchy in the data causes biases in random effects inferences and even biases the global fixed effects if missing values occur in the explanatory variables. A further finding was that incorporating dummies for the clusters in the imputation model improves the inferences for the global fixed effects but the estimated variances of the random effects can still be biased. Still, this work was limited to random intercepts and did not explain the results analytically.

Recently, several theoretical articles, comparing imputation methods in a multilevel setting, have appeared. Drechsler

(2015) theoretically explained the bias found in the simulations of van Buuren (2011) and illustrated that the bias depends on the cluster size, the missing data rate, and the intra-class correlation (ICC), which, in random intercepts models, is the proportion of variance between clusters relative to the total variance. Like van Buuren (2011), he only focused on random intercept models. Lüdtke et al. (2017) again only focused on random intercept models. They compared a single level imputation (which ignores the clustering of the data), a cluster-specific fixed-effects imputation (incorporating cluster-specific intercepts), and a multilevel imputation with respect to the bias in the intra-class correlation. They derived their results analytically and included a simulation study. Generally, they favored the multilevel imputation, but in some settings the single level imputation performed acceptable as well. The dummy imputation could be appropriate when the clusters and ICC are large and when the focus is on the regression coefficients only. The first paper to also consider random slopes was published by Grund et al. (2016). The authors evaluated the performance of two multilevel imputation strategies and listwise deletion under various settings. They found that the multilevel imputation methods worked well, as long as the missing data only occur in the dependent variable. If missings occur in the covariates, then random effects variances would be biased, an issue we will discuss later. The authors did not consider the dummy variable approach as an alternative to the multilevel imputation model.

Enders et al. (2016) mainly compared joint modeling (imputing all variables in one step) and sequential regression (imputing the variables step by step) in a setting of random intercepts and random slopes. Besides these imputation techniques, they also evaluated the performance of single-level imputation and including dummy variables for cluster-specific intercepts, but not cluster-specific slopes. They found that joint modeling and sequential regression produced similar results in random intercepts models. Joint modeling performed better when contextual effects (cluster means, etc.) were incorporated into the analysis model, while the sequential regression approach performed best in random slopes settings. The poor performance of the dummy variable and joint modeling approach in the random slopes context is not surprising since, except for the sequential regression approach, the authors only considered models that ignore the cluster-specific slopes. Finally, Zhou et al. (2016) proposed an approach to impute a binary variable for rare events in a multilevel setting. The idea is to generate synthetic populations and then to draw plausible values for the missing values from the posterior predictive distribution based on these populations. Via simulation based on a random intercept model, they compared their approach with a single-level imputation, an imputation model with intercept dummies for strata and clusters, and a random intercepts

imputation model. Results indicated poor coverage rates for single-level imputation. The fixed- and random- effects imputation models and their approach worked mostly well with some shortcomings, and random slopes were not considered.

To summarize, while all these articles cover imputation strategies for hierarchical data, they are subject to three important limitations: They only consider random intercept models (Reiter et al., 2006; Andridge, 2011; van Buuren, 2011; Drechsler, 2015; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008; Lüdtke et al., 2017), they only rely on simulation studies to evaluate the impact of different imputation approaches (Reiter et al., 2006; van Buuren, 2011; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008), or they do not evaluate the cluster-specific fixed-effects imputation approach as an alternative to the multilevel imputation model (Grund et al., 2016). Our contribution to the literature is that we analytically generalize the findings regarding the cluster-specific fixed-effects imputation compared to the multilevel imputation model by considering a setting with (arbitrarily many) cluster-specific variable dummies. We also show which factors govern the potential bias from cluster-specific fixed-effects imputation.

Multilevel modeling

With hierarchical data, each individual belongs to one of J clusters. Assuming that individuals within the same cluster are relatively homogeneous, it makes sense to extend the standard linear regression model to account for this. For example, school classes can be homogeneous if the school district lies in an area with many pupils from a specific socio-economic group. As the literature has identified, the socio-economic background tends to be influential on many educational issues (American Psychological Association, 2017), and analyses about the students' educational abilities have to take this homogeneity into account. The multilevel model (or *linear mixed model* as it is often referred to in statistics) is an extension of the linear model and has been a common analysis model for hierarchical data for many years (see, for example, Hedeker and Gibbons 1997 or Verbeke and Molenberghs 2009). A multilevel model incorporates cluster level *random effects* in addition to the global *fixed effects* to take the data hierarchy into account. The general multilevel model is given by:

$$\begin{aligned} y_{ij} &= X_{ij}\beta + Z_{ij}\gamma_j + \varepsilon_{ij}, \\ \gamma_j &\sim N(0, \Sigma), \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \end{aligned} \quad (1)$$

where y_{ij} is the value of the target variable Y for individual $i = 1, \dots, n_j$ in cluster $j = 1, \dots, J$, with n_j being

the size of cluster j . X_{ij} is a $(1 \times P)$ vector containing the variables for which a constant effect across all clusters is assumed (generally this will include a 1-column for the intercept). β is the $(P \times 1)$ vector containing the global fixed effects. Z_{ij} is a $(1 \times K)$ vector containing the variables for which it is assumed that the effects vary between the clusters. Often Z is a subset of X , meaning that a variable can either have only a global fixed effect or both a global fixed effect and random effects, but will never be modeled as having random effects only. γ_j is a $(K \times 1)$ vector containing the cluster-specific random effect(s) for cluster j . They allow the effect(s) of Z to vary between the clusters and are assumed to follow a multivariate normal distribution with zero-mean and covariance matrix Σ . This modeling strategy implicitly assumes that the observed clusters represent a random selection from a larger population of clusters. The assumption is met if 1,000 schools in the U.S. are sampled from the existing 100,000+ schools, but when characteristics are measured on all 50 U.S. states, including random effects for the states, it is not appropriate, as the states are the basic population and not a sample from it. For later use, we define γ to be the $J \times K$ matrix containing all random effects $\gamma = (\gamma'_1, \dots, \gamma'_J)'$. Finally, ε_{ij} is the error term and σ_ε^2 its variance.

To give an example in which situation the multilevel modeling approach could be used in educational research, consider the following model that analyzes the relationship between the score in a math test in year 1 and in year 3 of schooling:

$$\begin{aligned} scoreyear3_{ij} &= \beta_0 + scoreyear1_{ij} \cdot \beta_1 \\ &\quad + \gamma_{0j} + scoreyear1_{ij} \cdot \gamma_{1j} + \varepsilon_{ij} \end{aligned} \quad (2)$$

This modeling strategy would imply that there is a global average score β_0 (say 10) that students have in year 3 if their score in year 1 was 0. For each additional point scored in year 1, the expected score in year 3 increases by β_1 (say 0.8) points, on average. Now, for each cluster, these effects are assumed to vary randomly around the global effects. For example, it could be the case that in school 27 the expected average score is higher (say 11.5, implying $\gamma_{0,27} = 1.5$) but the effect of the test in year 1 is lower (say 0.6, implying that $\gamma_{1,27} = -0.2$).

Imputation models

Imputation methods based on the multiple imputation approach generally consist of two steps: First, a set of model parameters is drawn from their posterior distributions given the data. In the second step, missing values are replaced by repeated draws from the specified distribution given the parameters drawn from step one. This section describes these two steps for the two imputation models

to be compared: the *cluster-specific fixed-effects imputation* and the *multilevel imputation* model.

Cluster-specific fixed-effects imputation

The easiest way to extend the standard linear (multiple) imputation procedures to account for the hierarchy in the data is to incorporate individual fixed effects for each cluster. In this case, the parametric model is given by:

$$y_{ij} = X_{ij}\beta + Z_{ij}\gamma_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (3)$$

The only (yet crucial) difference to Eq. 1 is that γ_j is no longer assumed to be a realization from a normal distribution, but rather assumed to be fixed. In practice, this implies that a dummy variable for each cluster is included in the model and each variable in Z is interacted with this dummy. Let $I_j = I(y_{ij} \in cluster_j)$ be the indicator function that equals 1 if y_{ij} belongs to cluster j , and equals zero otherwise. The model to be estimated is given by:

$$y_{ij} = \sum_{p=1}^{P-T} X_{ijp}\beta_p + \sum_{j=1}^{J-1} \sum_{k=1}^K Z_{ijk}I_j\gamma_{jk} + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right), \quad (4)$$

where $p = 1, \dots, P$ is the index for the P variables in X and $k = 1, \dots, K$ is the index for the K variables contained in Z . Without loss of generality, we assume that X is sorted so that those T variables in X , that are also included in Z , are included in the last T columns of X . These variables need to be dropped (in addition to the reference categories for the dummy variables) to keep the model identified.

Since this is a standard linear regression model, with the usual assumption of uninformative priors (Bartlett et al., 2015), the draws for the first step of the imputation come from the following posterior distributions:

$$\tilde{\sigma}_\varepsilon^2 \sim \chi^{-2}\left(n^{obs} - d, \left[n^{obs} - d\right] \cdot \hat{\sigma}_\varepsilon^2\right),$$

$$\tilde{\delta} \sim N\left(\hat{\delta}, \left[V^{obs'}V^{obs}\right]^{-1} \cdot \tilde{\sigma}_\varepsilon^2\right), \quad (5)$$

where χ^{-2} is an inverse Chi-squared distribution, n^{obs} is the number of individuals over all clusters for which the outcome Y is observed, and $d = (J - 1) \cdot K + P - T$ is the number of coefficients that need to be estimated. $\delta = \{\gamma_{11}, \dots, \gamma_{(J-1)K}, \beta_1, \dots, \beta_{P-T}\}$ is the collection of parameters to be estimated and $V = \{Z_1I_1, \dots, Z_1I_{J-1}, Z_2I_1, \dots, Z_KI_{J-1}, X_1, \dots, X_{P-T}\}$ is the matrix of explanatory variables. V^{obs} is the subset of V containing those observations for which the outcome Y is observed. Note that we assume that all explanatory variables are fully observed or that missing values in these variables have been imputed in previous steps, as missing values in

an explanatory variable can cause biases in some parameter estimates (Grund et al., 2016). Lastly, $\hat{\sigma}_\varepsilon^2$ and $\hat{\delta}$ are the ordinary least squares estimates for σ_ε^2 and δ .

In the second step, missing values are imputed by randomly drawing values from

$$Y^{imp} \sim N(V^{imp}\tilde{\delta}, \tilde{\sigma}_\varepsilon^2), \quad (6)$$

where Y^{imp} and V^{imp} denote the subset of Y and V for which Y is missing.

Multilevel imputation

Since the posterior distribution of the parameters of the multilevel model is not available in closed form, a Gibbs sampler is required for the first step of the imputation (see for example Gelman and Hill (2006) for details). Assuming uninformative priors, draws from the following conditional models need to be iterated until convergence (for readability we use $|\cdot$ to indicate conditioning on all other parameters and the data at each step of the Gibbs sampler):

The global fixed effects for the imputation model are drawn from the normal posterior distribution:

$$\tilde{\beta}|\cdot \sim N(\beta^*, \Sigma^*) \quad \text{with}$$

$$\beta^* = \left(X^{obs'}X^{obs}\right)^{-1}X^{obs'}\left(y^{obs} - Z^{obs}\tilde{\gamma}\right)$$

$$\Sigma^* = \tilde{\sigma}_\varepsilon^2 \cdot \left(X^{obs'}X^{obs}\right)^{-1} \quad (7)$$

The residual variance is based on the posterior χ^2 distribution with $n^{obs} - 1$ degrees of freedom

$$\tilde{\sigma}_\varepsilon|\cdot \sim \sqrt{\frac{\sum_{j=1}^J \sum_{i=1}^{n_j^{obs}} \left(y_{ij}^{obs} - X_{ij}^{obs}\tilde{\beta} - Z_{ij}^{obs}\tilde{\gamma}_j\right)^2}{\chi_{n^{obs}-1}^2(n^{obs} - 1)}} \quad (8)$$

The variance of the random effects is drawn from the posterior Wishart distribution with $J + K$ degrees of freedom

$$\tilde{\Sigma}|\cdot \sim Wishart(\Sigma^*)_{J+K}^{-1} \quad \text{with}$$

$$\Sigma^* = \left(\tilde{\gamma}'\tilde{\gamma} + S_p\right)^{-1}, \quad (9)$$

where $S_p = K \cdot \hat{\Sigma}^{obs}$ is the prior for the random effects variance and $\hat{\Sigma}^{obs}$ is the estimated random effects variance based on the observed data.

The cluster-specific random effects are (multivariate) normally distributed

$$\tilde{\gamma}_j|\cdot \sim N(\gamma_j^*, \Omega_j)$$

$$\gamma_j^* = \left(Z_j^{obs'}Z_j^{obs} + \tilde{\sigma}_\varepsilon^2 \cdot \tilde{\Sigma}^{-1}\right)^{-1}Z_j^{obs'}\left(y_j^{obs} - X_j^{obs}\tilde{\beta}\right)$$

$$\Omega_j = \tilde{\sigma}_\varepsilon^2 \cdot \left(Z_j^{obs'}Z_j^{obs} + \tilde{\sigma}_\varepsilon^2 \cdot \tilde{\Sigma}^{-1}\right)^{-1} \quad (10)$$

In the second step, missing values in Y are imputed by drawing from:

$$y_{ij} \sim N\left(X_{ij}\tilde{\beta} + Z_j\tilde{\gamma}_j, \tilde{\sigma}_\varepsilon^2\right). \quad (11)$$

Theoretical juxtaposition of the two imputation models

Both imputation models have an important common feature: they allow one to incorporate cluster-specific effects. The main difference is that cluster-specific fixed-effects imputation assumes that the cluster effects are fixed quantities, whereas in multilevel models it is assumed that the cluster effects are random deviations from the global effect and these deviations follow a known distribution.

Including many dummy variables for the cluster-specific fixed-effects imputation can result in a large amount of parameters to be estimated (cf. Enders et al., 2016). On the other hand, one drawback of the multilevel imputation is its computational complexity resulting in relatively long run times and the task to monitor convergence of the imputation runs. It is well known (see, for example, Wooldridge 2010) that both models provide consistent estimates of the global fixed effects in a multilevel analysis model. However, as illustrated by Reiter et al. (2006) and Andridge (2011), the estimated variances of these global fixed-effects estimates will be biased after a cluster-specific fixed-effects imputation.

Furthermore, because the cluster-specific effects are modeled differently within the imputation, we also expect that the inferences of the random effects will be affected in the analysis model. Since the variance components are often of major interest in multilevel modeling, we will focus on the impact on the estimated covariance matrix of the random effects.

Directly quantifying the impact is difficult since the distribution of the random effects cannot be obtained in closed form. Thus, we follow the approach of Drechsler (2015) and compare the covariance matrix of the cluster-specific effects conditioning on all other parameters in the model. Since for the cluster-specific fixed-effects approach the conditional cluster-specific effects in one cluster are independent of the other clusters this conditional covariance matrix can be computed based solely on the information from the cluster. For cluster j , the matrix is given by (see Appendix A for details):

$$\begin{aligned} \text{Var}\left(\gamma_j^{fix}|\beta, V^{obs}\right) &= \sigma_\varepsilon^2 \cdot \left(Z_j^{obs'} Z_j^{obs}\right)^{-1} \\ &= \left(1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs}\right)^{-1}, \end{aligned} \quad (12)$$

where $\gamma_j^{fix} = \{\gamma_{1j}, \dots, \gamma_{Kj}\}'$ is the collection of cluster-specific fixed effects, $\beta = \{\beta_1, \dots, \beta_{P-T}\}$ is the collection

of global fixed effects, V^{obs} is the observed data, and Z_j^{obs} is the subset of records in Z_j for which Y is observed, where Z_j contains those variables in cluster j for which cluster-specific effects are assumed (in the example above Z_j is a matrix with a column of 1s for the intercept and the score of students in year 1 from class j). As noted above, the same conditional covariance matrix for the multilevel model is given by (Goldstein 2011 p. 69)

$$\begin{aligned} \text{Var}\left(\gamma_j^{multi}|\dots, V^{obs}\right) &= \sigma_\varepsilon^2 \cdot \left(Z_j^{obs'} Z_j^{obs} + \sigma_\varepsilon^2 \cdot \Sigma^{-1}\right)^{-1} \\ &= \left(1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs} + \Sigma^{-1}\right)^{-1}. \end{aligned} \quad (13)$$

The analytic comparison of Eqs. 12 and 13 is the main part of this section and key to this article. In the appendix, these equations are compared in detail regarding their Loewner-ordering (a mathematical concept to compare matrices), their additive and multiplicative difference, and their representations as ellipsoids (a multidimensional generalization of two-dimensional ellipses). Here we want to limit ourselves to the major findings. The first major finding (see Appendix B for details and proofs): $\text{Var}(\gamma_j^{fix}|\cdot)$ is Loewner larger than $\text{Var}(\gamma_j^{multi}|\cdot)$ and therefore the variances of the estimated random effects are always larger for the cluster-specific fixed-effects imputation. Assuming a correctly specified analysis model, this implies that after cluster-specific fixed-effects imputation, the estimated variances on the second level of the multilevel analysis model will always have a positive bias. The second major finding (see Appendix C): The multiplicative difference between the two variances (14) allows one to draw many conclusions regarding the causes of bias induced by the fixed-effects imputation:

$$\begin{aligned} \text{Var}\left(\gamma_j^{fix}|\cdot\right) &= \left(I + \left[Z_j^{obs'} Z_j^{obs}\right]^{-1} \cdot \sigma_\varepsilon^2 \cdot \Sigma^{-1}\right) \\ &\quad \cdot \text{Var}\left(\gamma_j^{multi}|\cdot\right) \end{aligned} \quad (14)$$

On the one hand, the difference depends on the ratio of the two variance components σ_ε^2 and Σ . Higher random effects variances in Σ will decrease the bias, whereas higher residual variances σ_ε^2 will increase it. Intuitively this makes sense. If the residual variance σ_ε^2 (i.e., the variance on the individual level) is small relative to the cluster level variance Σ , this implies that all the variation is between the clusters and thus the multilevel model coincides with the cluster-specific fixed-effects model. Both imputation models will lead to similar results in this case. However, if the individual level variance is large relative to the cluster level variance, results based on a cluster-specific fixed-effects analysis model will differ from the results obtained from a multilevel analysis model and we would expect to see a similar effect if cluster-specific fixed-effects and multilevel models are used at the imputation stage.

Besides the ratio of the two variance components, the difference also depends on the matrix of explanatory variables in cluster j . Under rather general conditions, the difference decreases with increasing cluster size since the main diagonal elements of $(Z_j^{obs'} Z_j^{obs})^{-1}$ decrease as n_j increases (see Appendix D). Again, this is plausible, since the shrinkage effect of the multilevel model generally decreases with increasing cluster size and thus the differences between the two models also decreases with the size of the cluster. An implication that is easily overseen is that the difference will implicitly also depend on the missing data mechanism since Z_j^{obs} only contains those records for which Y is observed. If, for example, the missingness in Y is positively correlated with Z , i.e., the probability for Y to be missing is higher for larger Z , the matrix $Z_j^{obs'} Z_j^{obs}$ will look different than if the missingness is negatively correlated with Z . We will address this issue in the next section. We also note that Eq. 14 reveals that the bias does not depend on the number of available clusters since the number of clusters J does not appear in the equation.

The third major finding: The ellipsoid of the random effects after *cluster-specific fixed-effects imputation* always fully encloses the *multilevel imputation-ellipsoid* (see Appendix E). One interpretation is that the confidence region for the joint distribution of the conditional parameters for γ_j^{fix} fully encloses the confidence region for γ_j^{multi} for any significance level α . This allows us to make a more general statement compared to the first finding: the set of random effects (inspected jointly) will vary more in every possible direction (regardless of their covariance) after cluster-specific fixed-effects imputation. Thus, we would generally overestimate the variability on the second level of our multilevel model. This directly implies that the "classical" intra-class correlation $ICC = \sigma_0^2 / (\sigma_0^2 + \sigma_\epsilon^2)$, with σ_0^2 being the variance on the second level, will be positively biased in a random intercepts setting (the fraction increases as σ_0^2 increases while σ_ϵ^2 remains constant).

Simulation study

To evaluate whether the identified differences between the two models also lead to substantial bias in the inferences obtained from the imputed dataset, we run extensive simulation studies in R (R Core Team, 2016). The simulations (repeated 1000 times) consist of four steps:

1. Data generation
2. Inducement of nonresponse
3. Multiple ($M = 50$) imputation based on both the cluster-specific fixed-effects and multilevel imputation models described above
4. Running a multilevel analysis model on the imputed dataset

In the following, we will describe each step in detail.

Data generation

To limit the number of parameters that need to be evaluated, we assume the model of interest has, besides the random intercepts, just *one* random slope variable. We do not expect any further insights from the inclusion of further random coefficients.

For the simulation, we assume that the analysis model is correctly specified, i.e., the analysis model matches the data generating process. Of course, this assumption is often not met in practice; however, it is moot to discuss potential biases from imputation if the analysis model would already be biased in the absence of any missing data. For simplicity, we only include two explanatory global fixed-effects variables— W_1 varying at the individual level (e.g., the test score in year 1) and W_2 varying at the cluster level (e.g., the teachers age)—in our random coefficients analysis model. These two variables were generated according to the following models:

$$W_1 \sim N(1, 2 \cdot I_n)$$

$$W_2 \sim N(3, 1.5 \cdot I_J), \tag{15}$$

where I_n and I_J denote the identity matrices (a matrix with 1s on the main diagonal and 0s elsewhere) of dimension n and J , where n is the number of individuals and J is the number of clusters. In other words, we have n independent draws from a normal distribution with mean 1 and standard deviation 2 and J independent draws from a normal distribution with mean 3 and standard deviation 1.5. Our random coefficient model is given as:

$$Y = X\beta + Z\gamma + \epsilon, \quad \text{with } \epsilon \sim N(0, I_n \cdot \sigma_\epsilon^2), \tag{16}$$

where $X = \{1, W_1, W_2^*, W_1 \cdot W_2^*\}$, $Z = \{1, W_1\}$, and W_2^* is the cluster level variable W_2 "blown-up" to have the same length as the other variables by repeating each entry j n_j times, where n_j is the cluster size for cluster j and $j = \{1, \dots, J\}$. So the model has an intercept, two fixed-effects covariates, and their interaction as global fixed-effects variables in the model. Besides the fixed-effects variables, the model contains a random intercept and a random slope variable. The values of the global fixed effects are set to $\beta = \{2, 1, 1.5, -0.3\}$ and the random effects are generated as

$$\gamma \sim N(0, \Sigma) \quad \text{with}$$

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.7 & -0.3 \\ -0.3 & 0.8 \end{pmatrix} \tag{17}$$

We keep the cluster sizes equal for all clusters, but alter them across different simulation settings between 15, 25, and 50. The number of clusters is fixed at 30 and is not altered further as the number of clusters does not affect the bias (see

previous section). We run the simulation for different values of the residual variance σ_ε (1.0, 1.5, and 2.0), allowing us to examine the bias under different intra-class correlations. Furthermore, because the missing data mechanism (described in detail below) influences the bias, the simulation results are presented under five different models for the nonresponse.

The nonresponse model

Step two in the simulation design is the inducement of missing values. In our simulation, the missingness is limited to Y , and the missingness mechanism is modeled based on a logistic function of W_1 . Since we identified the missingness mechanism as influential for the amount of bias, we need a model that allows for some flexibility regarding the influence of W_1 on the probability of Y to be missing. We decided to use the following model:

$$P(Y_{ij} = \text{NA} | \tilde{W}_{1ij}, s) = MR \cdot (1 - s) + 2 \cdot MR \cdot s \cdot \text{logit}(\tilde{W}_{1ij})^{-1} \tag{18}$$

where MR denotes the desired missing rate, which we fix at 0.5. \tilde{W}_{1ij} is the standardized version of W_{1ij} , i.e., $\tilde{W}_{1ij} = (W_{1ij} - \bar{W}_1) / \sqrt{\text{var}(W_{1ij})}$. The parameter s governs the influence of W_1 on the probability of Y to be missing. Figure 1 illustrates the missing data probability functions for different settings of s . Using this model has several implications:

- To obtain a valid probability model, the range of s needs to be bounded by $\{\max(-1, [1 - 1/MR]), \min(1, [1 - MR]/MR)\}$. As we set $MR = 0.5$, s is bounded by $\{-1, 1\}$.
- $s = 0$ implies Missing Completely At Random (MCAR, see Rubin 1976).
- $s > 0$ ($s < 0$) implies a positive (negative) correlation between x and the probability to be missing and thus Missing At Random (MAR, see Rubin 1976).
- Larger values of $|s|$ imply a stronger influence of W_1 on the probability of Y to be missing.
- If \tilde{W}_1 is symmetrically distributed around 0, the expected missing rate over all records in a dataset is equal to MR .
- Records with \tilde{W}_1 values close to 0 will be missing with a probability equal to MR .
- The record with the smallest (resp. largest) possible W_1 value will have a probability for Y to be missing close to $(1 - s) \cdot MR$ (resp. $(1 + s) \cdot MR$).

In our simulations, we alter s within $\{-1, 0.5, 0, 0.5, 1\}$ to evaluate the impact of the missing data mechanism. Whenever less than six observed records remain in one of the

clusters, the missing data generation is repeated for this cluster to ensure numerical stability.

Parameters of interest

As discussed above, we assume that the analysis model of interest is a random slopes model that is congenial to the data generating process. Point estimates of the global regression parameters β should not be biased by a cluster-specific fixed-effects imputation procedure, so we do not focus on them. Instead, we look at the variances of the global fixed-effects and the random-effects variances σ_0^2 and σ_1^2 , often reported in educational research to evaluate how much of the total variance in the outcome variable is explained by the cluster level units. Both imputation methods are programmed using own code following the description in the section “Imputation models”. The functions for the multi-level imputation will be incorporated in the R package `hmi` by Speidel et al. (2017) in the future. All parameter estimations for the multilevel analysis model are computed using the function `lmer` from the R-package `lme4` by Bates et al. (2016).

Results of the simulation study

We discuss the impacts on the random effects first before describing the implications for the variances of the fixed effects. We only present results for the cluster-specific fixed-effects imputation. Results for the original data (before values were deleted) and for imputation based on the multilevel model did not show any significant bias and we omit them for brevity. In order to make the differences in the estimations $\hat{\theta}_{run}$, $run = 1, \dots, 1000$ for $\theta = \{\sigma_0^2, \sigma_1^2\}$ easily comparable, we look at the *empirical relative bias*:

$$\frac{\hat{\theta}_{run} - \theta}{|\theta|}, \theta \neq 0. \tag{19}$$

If, for example, the true value is 0.7 and the estimate 0.71, the empirical relative bias is $(0.71 - 0.7)/|0.7| \approx 0.014$, which is an overestimation of 1.4%. An unbiased method has an empirical relative bias of 0. As the simulations work empirically, not even the estimates on the original data will yield an empirical relative bias of exactly 0. Therefore, a small relative empirical bias is tolerable. As a rough guideline, we refer to Grund et al. (2016). They consider relative biases of $\pm 5\%$ for global fixed effects and $\pm 30\%$ for variance parameters to be noteworthy.

Implications for the random effects

Figures 2 and 3 show the relative empirical biases for the random intercepts and random slopes variances for all combinations of the cluster size, residual variance, and missing

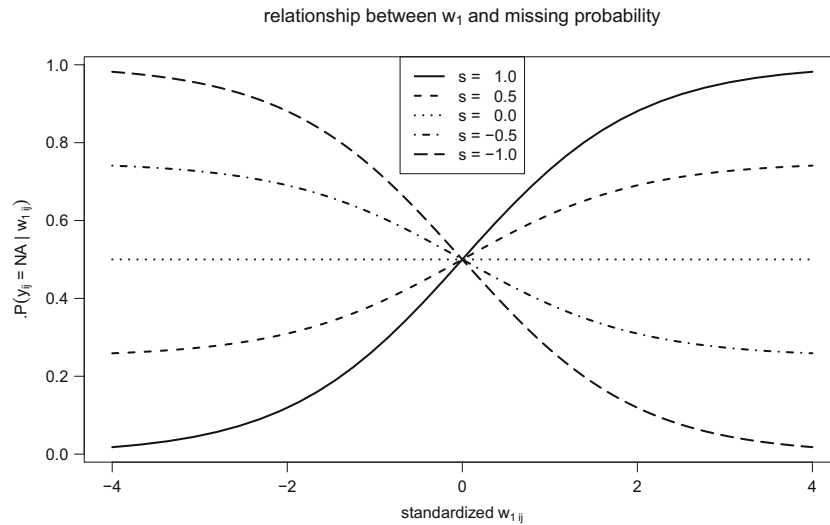


Fig. 1 Illustration of missing data probabilities for different settings of s (=relationship between w_1 and missing data probability) from strong positive ($s = 1.0$) over missing completely at random (MCAR; $s = 0.0$) to strong negative ($s = -1.0$)

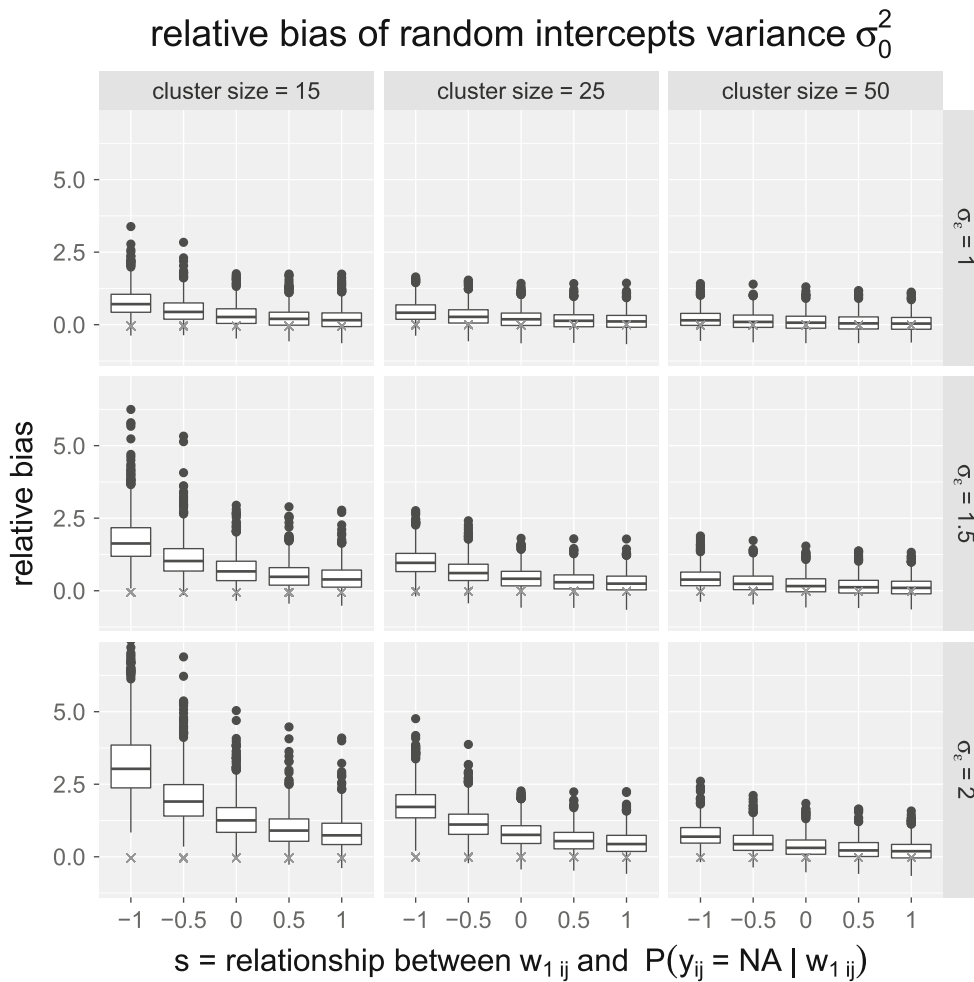


Fig. 2 Relative bias for the estimated variance of the random intercept. The cross marks the median empirical bias of the estimates on the original data as a reference. 10 points (out of 45k) larger than 7 are not shown for readability of the figure

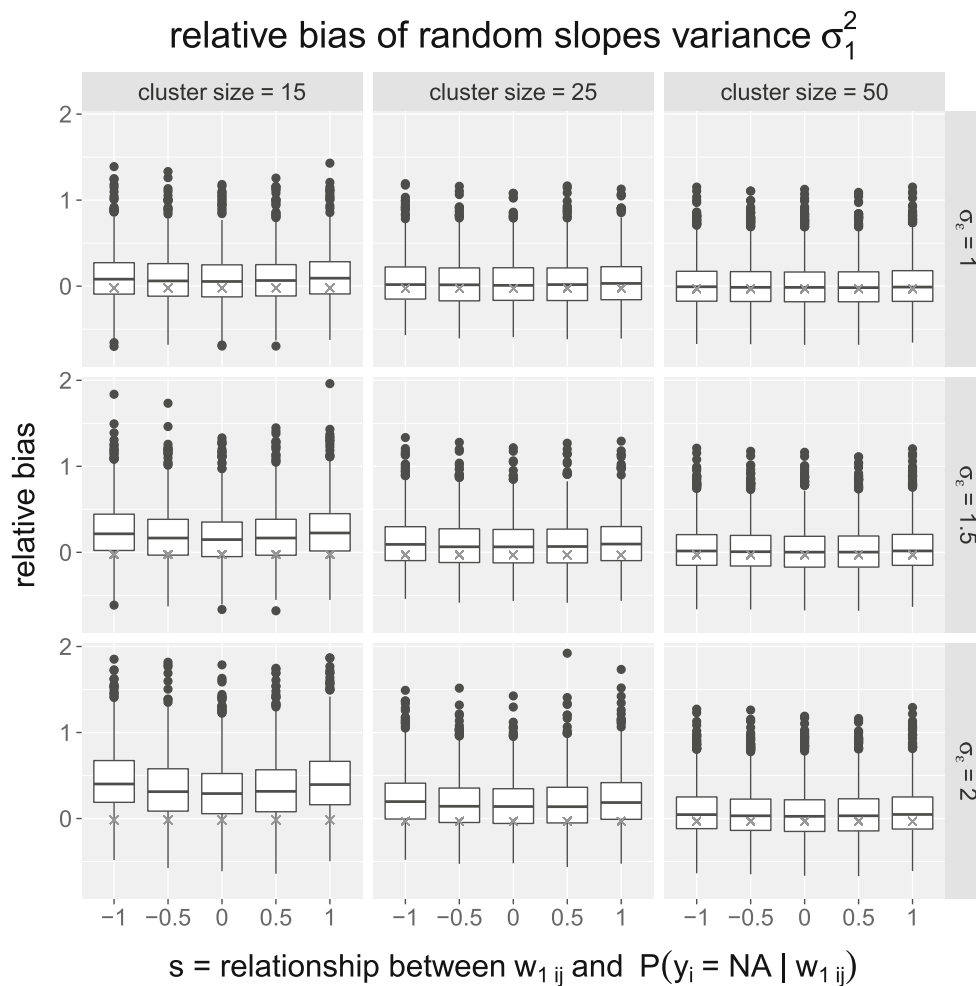


Fig. 3 Relative bias for the estimated variance of the random slope. The cross marks the median empirical bias of the estimates on the original data as a reference. 3 points (out of 45k) larger than 2 are not shown for readability of the figure

data mechanism. A boxplot centered around 0 indicates empirical unbiasedness.

In most settings, the random effects variances are overestimated. In some settings they are (practically) unbiased, but never underestimated. The amount of bias decreases with increasing cluster size, but increases with increasing residual variance. These results are in line with our derivations in the previous section. The bias for the random intercepts is generally larger than the bias for the random slopes (the median relative bias of the random intercept is almost three for $s = -1$, $\sigma_\varepsilon = 2$, and cluster size equal to 15, whereas the median relative bias of the random slopes never exceeds 0.5). We also see that the bias depends on the missing data mechanism. We find a decreasing bias with increasing s for the intercepts and a U-shaped effect for the slopes. It is difficult to explain the process behind these results in general because the bias is governed by distributional properties of Z_{obs} , the random effect variables of those individuals with

an observed target variable value (see Eq. 14). We provide some explanations for the observed relationship between the missing data mechanism and the bias for our specific setup in Appendix F.

Implications for the global fixed effects

We do not expect to see any bias in the point estimates of the global fixed effects since both the cluster-specific fixed-effects imputation model and the multilevel imputation model provide unbiased point estimates of the true population parameters. This was confirmed in our simulation study (results not shown for brevity). However, as Reiter et al. (2006) and Andridge (2011) point out, the variances of the global fixed effects should be overestimated if the cluster-specific fixed-effects imputation approach is used. Our simulation study also confirmed this finding. Figure 4 contains variance ratios for all global fixed effects for all

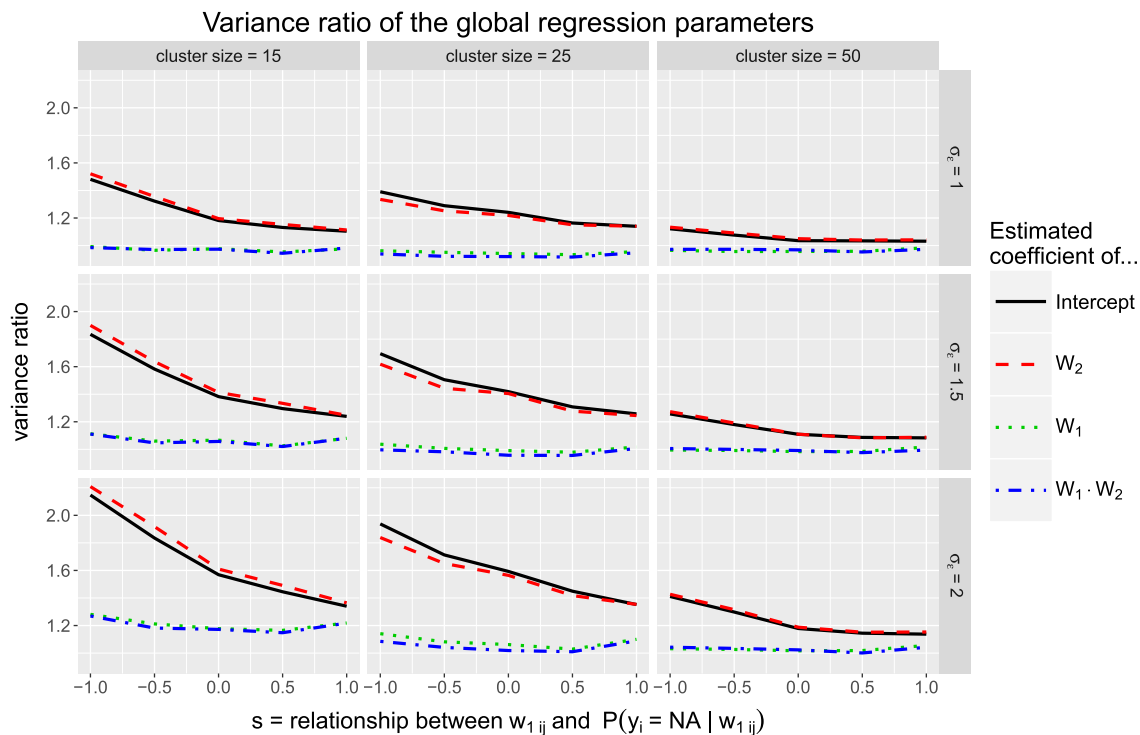


Fig. 4 The variance ratio of the global regression parameters (=median of estimated variances of $\hat{\beta}$ divided by the empirical variance of $\hat{\beta}$)

simulation setups. The variance ratios are computed by dividing the median estimated variance by the true variance of the point estimates across the 1,000 simulation runs. Most of the ratios are greater than 1 indicating that the variance is generally overestimated leading to conservative point estimates and an increased chance of type II errors. The few cases in which the variance ratios are less than 1 seem to be artifacts, since the variance ratios for the original data before deletion (not reported) are even smaller in these cases, indicating a general bias in the analysis procedure. However, beyond confirming results previously discussed in the literature, the figure also illustrates that there is a close relationship between the biases in the random effects variances and the biases in the global fixed-effects variances. As with the random effects, the biases in the variance ratios decrease with increasing cluster size and increase with increasing residual variance. As discussed above, these results are to be expected as the multilevel imputation and cluster-specific fixed-effects imputation become more similar with increasing cluster size and decreasing residual variance. The effect of the nonresponse mechanism s on the bias needs some further explanations. Note that the negative relationship between s and the bias for the regression coefficients of the intercept and W_2 follows the relationship found for the random intercept variances, whereas the U-shaped relationship for the regression coefficients of W_1

and the interaction between W_1 and W_2 mimics the relationship found for the random slopes variances. This can be explained if we note that we can express the random slopes model in Eq. (16) in a different way:

$$Y_{ij} = \alpha_j + \beta_j W_{1ij} + \varepsilon_{ij}, \quad \text{with } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N \left(\begin{bmatrix} \gamma_0^\alpha + \gamma_1^\alpha W_{2j} \\ \gamma_0^\beta + \gamma_1^\beta W_{2j} \end{bmatrix}, \Sigma \right). \quad (20)$$

Rewriting the model like this is helpful because it illustrates that there is a relationship between the random intercepts α_j and the coefficients γ_0^α and γ_1^α , and likewise a relationship between the random slopes β_j and γ_0^β and γ_1^β . Relating this notation to the notation in Eq. 16, γ_0^α and γ_1^α are the regression coefficients for the intercept and W_2 , while γ_0^β and γ_1^β are the coefficients for W_1 and the interaction between W_1 and W_2 . This explains why the biases for the variance of the coefficients of the intercept and W_2 follow a similar pattern as the bias of the random intercept variance. Likewise, we better understand the relationship between the biases for the variance of the coefficients of W_1 and the interaction term, and the bias of the random slopes variance. To our knowledge, this topic has not been addressed in the literature so far and determining the exact relationship between the two effects analytically would be an interesting topic for future research.

Table 1 Estimates of the variance parameters for National Educational Panel Study data

parameter	available case analysis	fixed-effects imp.	multilevel imp.
σ_0^2	0.0228	0.0310	0.0232
σ_1^2	0.0869	0.1167	0.0844
σ_{01}	-0.0441	-0.0596	-0.0449
σ_ε^2	0.1056	0.1058	0.1057

Real data application

In this section, we evaluate whether our theoretical and simulation-based findings are relevant in an applied setting. An appropriate field of research where random effect variances are of particular interest is the evaluation of teacher effectiveness (see e.g., Lenkeit 2012, Nye et al., 2004, or more generally McCaffrey et al., 2004b). Thus, we use data from the Starting Cohort 3 of the National Educational Panel Study (Blossfeld et al., 2011). The NEPS, run by the Leibniz Institute for Educational Trajectories, is an extensive study in Germany that aims to measure the reasons and impacts of educational decisions over the entire life course. To achieve this goal, surveys are conducted in a multi cohort sequence design in which six different cohorts are followed for several years. The cohorts are selected to cover the entire life span starting with an infant cohort, a kindergarten cohort, a cohort of pupils in elementary school, etc. The final cohort is an adult cohort that represents adults aged 23 to 64 by the time of the first interview. The six starting cohorts were recruited between 2009 and 2012 containing more than 60,000 target persons. 5th grade students comprise the Starting Cohort

3. Many performance related items have been administered to these students, including two math competence tests in years 2010 and 2012. Because the aim of this real data application is not to draw conclusions from sophisticated educational analyses, but to show the impact of the imputation methods, we consider a simple model for students' achievements in math competence tests:

$$\mathit{math7}_{ij} = \beta_0 + \gamma_{0j} + \mathit{math5}_{ij} \cdot (\beta_1 + \gamma_{1j}) + \varepsilon_{ij}, \quad (21)$$

where variable *math5* is the 5th grade test score and *math7* the test score in grade 7. We conditioned on those students who had the same teacher in both years and only missing values in *math7*. This resulted in $n = 630$ students overall and 29 students having missing values in *math7* (\rightarrow missing rate of 4.6%). We multiply ($M = 50$) imputed *math7* using both methods (the cluster-specific fixed-effects imputation and the multilevel imputation) and estimated the multilevel model from Eq. 21 with `lmer`. To isolate the teacher effect, several additional control variables would normally be included on both levels of the model in practice (see, for example, McCaffrey et al., 2004a). Since including more variables will not provide additional insights regarding the implications of the two imputation strategies, we keep the model simple for illustrative purposes.

Because we only focus on those observations for which missingness is limited to the dependent variable, we can use available case analysis as a benchmark, since available case methods will provide unbiased estimates in this case (assuming the model is correctly specified). The estimations of the random effects variances in Table 1 are in line with the findings from the theoretical section and the simulation study: after the cluster-specific fixed-effects imputation the variance estimates are recognizably higher while the estimates based

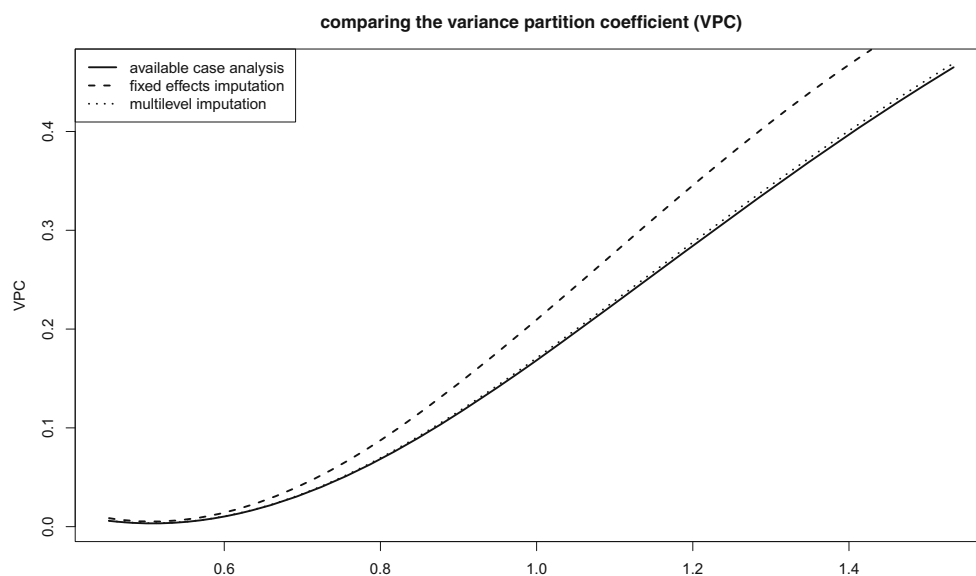
**Fig. 5** The variance partition coefficient (VPC) for the range of $x = \mathit{math5}$ in the NEPS data

Table 2 Point estimates and confidence interval (CI) properties of the global fixed effects for National Educational Panel Study data

parameter	point estimate	0.025%-quantile	0.975%-quantile	CI-length
β_0^{aca}	0.3830	0.3123	0.4537	0.1414
β_0^{fix}	0.3808	0.3027	0.4589	0.1562
β_0^{multi}	0.3823	0.3105	0.4541	0.1437
β_1^{aca}	0.2914	0.1563	0.4266	0.2703
β_1^{fix}	0.2967	0.1473	0.4461	0.2988
β_1^{multi}	0.2924	0.1550	0.4298	0.2748

on multilevel imputation are very close to the benchmark values based on available case analysis. The overestimation after cluster-specific fixed-effects imputation is substantial considering that less than 5% of the data were imputed. As mentioned above, a commonly computed measure for the impact of the clustering on the total variance is the intra-class correlation (ICC). For models with more than only random intercepts, the "classical" $ICC = \sigma_0^2 / (\sigma_0^2 + \sigma_\varepsilon^2)$ is no longer sufficient to summarize the contribution of the clusters to the total variance. Goldstein et al. (2002) proposed the *variance partition coefficient* (VPC). The VPC is a function of the predictor variable x and the variance components and shows the 'importance' of the clusters for different values of x :

$$vpc = \frac{\sigma_0^2 + 2 \cdot \sigma_{01} \cdot x + \sigma_1^2 \cdot x^2}{\sigma_0^2 + 2 \cdot \sigma_{01} \cdot x + \sigma_1^2 \cdot x^2 + \sigma_\varepsilon^2} \quad (22)$$

Figure 5 shows the VPCs based on the available cases, the data after the cluster-specific fixed-effects imputation and after the multilevel imputation. One can see that the clusters would be viewed as being more 'important' under cluster-specific fixed-effects imputation than under multilevel imputation.

Results for the global fixed effects are presented in Table 2. The point estimates are almost identical for all methods. Considering the uncertainty of the estimates as expressed by the 95% confidence intervals, the difference between the inferences based on the three different analysis strategies is small, since the confidence intervals overlap to a large extent. The last column of the table shows that the confidence intervals for the global fixed effects are larger for the cluster-specific fixed-effects imputation, which is also in line with theoretical expectations.

Conclusion

Contributing to the discussion about suitable imputation methods for hierarchical data, we present theoretical and

empirical evidence to the supposition that the cluster-specific fixed-effects imputation is highly likely to bias variance parameter estimates in a multilevel analysis model. A simulation showed that the bias can be severe. Even though the simulation study was limited to random intercepts and random slopes, the theory holds for any number of random effects variables (starting from only random intercepts models and ending with models where each variable is treated as random). Therefore, we generally advise using multilevel imputation models. Even if there are only random intercepts, including a random slope variable should do no harm ("Consider all coefficients as potentially varying", Gelman and Hill 2006, p. 549).

The high variance in the cluster-specific effects under a cluster-specific fixed-effects imputation also negatively affects the coverage rates of the global fixed effects and increases the probability of false conclusions. In a real data application, we showed that even for a small missing data rate (less than 5%) the results can substantially differ. This is a further reason to use multilevel models for imputation.

A shortcoming of the multilevel imputation is the relatively high runtime. While the cluster-specific fixed-effects imputation took between 0.15 and 4 s (median 0.28) in our simulation settings, the multilevel imputation took between 1.5 and 5.2 min (median 2.2). This can be a severe drawback if many variables need to be imputed in a dataset. A second technical shortcoming of the multilevel imputation is the need to monitor convergence, which is not needed for the cluster-specific fixed-effects imputation as all posterior distributions can be obtained in closed form.

As Eq. 14 showed, both imputation methods will produce similar (if not identical) results in three conditions: large cluster sizes, large differences between the clusters, and small residual variances. We cannot provide general thresholds for these parameters to be "large" or "small," but if the researcher sees one (or better yet, as our simulation showed: more) of these conditions met, s/he might consider using a cluster-specific fixed-effects imputation instead of a multilevel imputation model for convenience. This is especially relevant if there are many variables to impute and not enough time to conduct multilevel imputations.

In this paper, we limited our analysis to missing values in the target variable as a starting point. Still, it would be worthwhile to investigate the impacts of missing values in the covariates. Grund et al. (2016) conducted a simulation study for this scenario and found some results after multilevel imputation to be biased. With the presumed gold standard to be biased, analytic explanations are needed to elucidate this phenomenon. Further research on higher level models, comprising more than two levels, or cross-classified clusters would also be beneficial.

Acknowledgements This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Priority Programme “Education as a Lifelong Process” [SPP 1646] - DR 831/2-2. This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 - 5th Grade, doi:10.5157/NEPS:SC3:3.1.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

To simplify all of the following equations, we define $A := 1/\sigma_\epsilon^2 \cdot Z_j^{obs'} Z_j^{obs}$ and $B := \Sigma^{-1}$. This allows us to write:

$$\begin{aligned} \text{Var}(\gamma_j^{fix} | \cdot) &= (A)^{-1} \\ \text{Var}(\gamma_j^{multi} | \cdot) &= (A + B)^{-1}. \end{aligned} \tag{23}$$

Appendix A

Following Drechsler (2015) we write $\phi' = (\gamma', \beta')$ for the set of cluster-specific and global effects in Eq. 4. $R = (Z_{ind}, X)$ is the $n \times ([J - 1] \cdot K + P - T)$ matrix of regression variables, with Z_{ind} being the $n \times ([J - 1] \cdot K)$ matrix of intercept and slope dummies. Generally, for linear models it holds that $\phi | \mu, R, \sigma_\epsilon^2 \sim N(\mu, \Sigma = \sigma_\epsilon^2 \cdot [R'R]^{-1})$. Let $\mu' = (\mu'_1, \mu'_2)$ be partitioned so that μ_1 contains the expected values for γ and μ_2 contains the expected values for β . We partition Σ in a similar way so that Σ_{11} consists of a $(J - 1) \cdot K \times (J - 1) \cdot K$ matrix containing the covariance matrix of the cluster-specific effects. Likewise, Σ_{22} is the $(P - T) \times (P - T)$ dimensional matrix containing the covariance matrix of the global effects. With this partitioning we have

$$\begin{aligned} \begin{pmatrix} \gamma^{fix} | \mu, R, \sigma_\epsilon^2 \\ \beta^{fix} | \mu, R, \sigma_\epsilon^2 \end{pmatrix} &\sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \\ &= \sigma_\epsilon^2 \cdot \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}^{-1} \end{aligned} \tag{24}$$

Since the joint distribution of the effects is multivariate normal it holds that the conditional variance of the cluster-specific effects given the global effects is

$$\text{Var}(\gamma^{fix} | \mu, R, \sigma_\epsilon^2, \beta^{fix}) = \Sigma_{11} - \Sigma_{12}[\Sigma_{22}]^{-1}\Sigma_{21} \tag{25}$$

In order to simplify this equation we use Harville (1997, Corollary 8.5.12 p. 100) which shows that for a nonsingular matrix $1/\sigma_\epsilon^2 \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}$ and its inverse $\sigma_\epsilon^2 \cdot \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}^{-1} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, partitioned in the same way, it holds that $\Sigma_{11} = \sigma_\epsilon^2 \cdot [(R'R)_{11}]^{-1} + \Sigma_{12}[\Sigma_{22}]^{-1}\Sigma_{21}$. Replacing this expression of Σ_{11} in Eq. 25 yields $\text{Var}(\gamma^{fix} | V^{obs}, \beta^{fix}) = \sigma_\epsilon^2 \cdot [(R'R)_{11}]^{-1}$.

Multiplication rules for block matrices yield $\sigma_\epsilon^2 \cdot \{([Z_{ind}, X]'[Z_{ind}, X])_{11}\}^{-1} = \sigma_\epsilon^2 \cdot \{Z'_{ind} Z_{ind}\}^{-1}$. For a cluster j this means that $\text{Var}(\gamma_j^{fix} | V^{obs} \beta^{fix}) = \sigma_\epsilon^2 \cdot (Z'_j Z_j)^{-1}$.

Appendix B

Here we show that the conditional variance of each of the cluster-specific fixed effects in the cluster-specific fixed-effects imputation model is larger than the conditional variance of the corresponding random effect in the multi-level imputation model, i.e., $\text{diag}[\text{Var}(\gamma^{fix} | \cdot)] > \text{diag}[\text{Var}(\gamma^{multi} | \cdot)]$. To start our proof, we look at the additive disparity of $\text{Var}(\gamma^{fix} | \cdot)$ and $\text{Var}(\gamma^{multi} | \cdot)$.

$$\begin{aligned} \text{Var}(\gamma_j^{multi} | \cdot) + \Delta &= \text{Var}(\gamma_j^{fix} | \cdot) \Leftrightarrow \\ \Delta &= \text{Var}(\gamma_j^{fix} | \cdot) - \text{Var}(\gamma_j^{multi} | \cdot) \Leftrightarrow \\ \Delta &= A^{-1} - (A + B)^{-1} \end{aligned} \tag{26}$$

To show that $\text{diag}[\text{Var}(\gamma^{fix} | \cdot)] > \text{diag}[\text{Var}(\gamma^{multi} | \cdot)]$, we need to show that Δ is positive definite since the main diagonal elements of positive definite matrices are always positive (Harville 1997, corollary 14.2.13 p. 214).

According to the definition of the Loewner order a Hermitian matrix M_2 is Loewner larger than a Hermitian matrix M_1 ($M_2 >_L M_1$) if the difference $M_2 - M_1$ is positive definite. It also holds that if $M_2 >_L M_1$ then $M_1^{-1} >_L M_2^{-1}$ (see e.g. Siotani 1967 eq. 3 p. 246 or Horn and Johnson 1990 theorem 7.7.4 p. 471). So to show that Δ is positive definite we need to show that $A^{-1} >_L (A + B)^{-1}$ or equivalently that $A + B >_L A$. The last statement is trivially fulfilled because the difference $A + B - A = B$ is positive definite (because B is the inverse of the positive covariance matrix Σ and the inverse of a positive definite matrix is also positive definite Harville 1997, corollary 14.2.11 p. 214).

Appendix C

To further understand which factors influence the difference between $\text{Var}(\gamma^{fix} | \cdot)$ and $\text{Var}(\gamma^{multi} | \cdot)$ it is informative to identify the multiplicative factor H for which it holds that:

$$H \cdot \text{Var}(\gamma_j^{multi} | \cdot) = \text{Var}(\gamma_j^{fix} | \cdot). \tag{27}$$

Simple matrix manipulation reveals that

$$\begin{aligned}
 H(A + B)^{-1} &= (A)^{-1} \Leftrightarrow \\
 H(A + B)^{-1} \cdot (A + B) &= (A)^{-1}(A + B) \\
 H &= (A)^{-1}A + (A)^{-1}B \\
 H &= I + (A)^{-1}B \\
 H &= I + \left(1/\sigma_\epsilon^2 \cdot Z_j^{obs'} Z_j^{obs}\right)^{-1} \Sigma^{-1}
 \end{aligned}
 \tag{28}$$

Appendix D

If Z is a $n_0 \times p$ data matrix then observing additional $n_1 \times p$ data Z_{new} results in the $(n_0 + n_1) \times p$ data matrix $Z^* = \begin{pmatrix} Z \\ Z_{new} \end{pmatrix}$. According to the multiplication rules for block matrices (see e.g., Harville 1997 section 2.2) it holds that

$$Z^* Z^* = \begin{pmatrix} Z' & Z'_{new} \end{pmatrix} \begin{pmatrix} Z \\ Z_{new} \end{pmatrix} = Z'Z + Z'_{new}Z_{new}.
 \tag{29}$$

Given the results in Appendix B it follows that $(Z^* Z^*)^{-1} <_L (Z'Z)^{-1}$. This shows that $(Z_j^{obs'} Z_j^{obs})^{-1}$ decreases as n_j^{obs} increases.

Appendix E

The $(1 - \alpha)$ -confidence region for a $p \times p$ covariance matrix can be represented as a p -dimensional ellipsoid (see e.g., Press et al., 2007 or Scheffé 1999). An ellipsoid for a matrix Σ is the set of points $z \neq 0$ that fulfill the equation $z' \cdot \Sigma^{-1} \cdot z = c$ with c being a constant scalar. The value of c can be used to define the $(1 - \alpha)$ -confidence ellipsoid. Therefore we give c a subscript $\delta := 1 - \alpha$ and write c_δ . If c_δ increases the ellipsoid becomes larger (covers more area/volume). So a larger c_δ means a higher level of certainty. We will show that for any value of z the ellipsoid equation for $Var(\gamma^{multi}|\cdot)$ will give a higher critical value c_δ than for $Var(\gamma^{fix}|\cdot)$, i.e., $z' \cdot Var(\gamma^{multi}|\cdot)^{-1} \cdot z = c_\delta^{multi} > c_\delta^{fix} = z' \cdot Var(\gamma^{fix}|\cdot)^{-1} \cdot z \Leftrightarrow \delta^{multi} > \delta^{fix}$. To make this point clearer let us assume $\delta^{multi} = 0.99 > \delta^{fix} = 0.90$. This would imply that only 1% of the data drawn based on $Var(\gamma^{multi}|\cdot)$, but 10% of the data drawn based on $Var(\gamma^{fix}|\cdot)$, are expected to exceed the point z . So for any value of δ the *fix*-ellipsoid fully encloses the *multi*-ellipsoid. And if an ellipsoid E_2 , representing matrix M_2 , fully encloses another ellipsoid E_1 , representing matrix M_1 , one can say that values drawn from M_2 vary more in every possible direction than values drawn from M_1 . The proof for our case is simple. It uses the result from Appendix B: $Var(\gamma^{multi}|\cdot)^{-1} >_L Var(\gamma^{fix}|\cdot)^{-1}$ which is equivalent

to $z' \cdot Var(\gamma^{multi}|\cdot)^{-1} \cdot z > z' \cdot Var(\gamma^{fix}|\cdot)^{-1} \cdot z$ for any $z \neq 0$.

Appendix F

Here we want to provide explanations why the random intercepts variance bias is negatively correlated with s (= relationship of W_1 and $P(y = NA|W_1)$) and the random slopes bias shows a U-shaped pattern for our specific data setting. As mentioned in the analytical section, the properties of Z_j^{obs} (the observations of Z_j for which Y is observed after nonresponse was generated) influence the $(Z_j^{obs'} Z_j^{obs})^{-1}$ part of the multiplicative difference between $Var(\gamma^{multi}|\cdot)$ and $Var(\gamma^{fix}|\cdot)$ (see also Equations 27 and 28 in Appendix C).

In our random slopes setting $Z = (1, W_1)$. This implies that $(Z_j^{obs'} Z_j^{obs})^{-1}$ becomes

$$\begin{aligned}
 &\left(n_{obs\ j} \cdot \sum_{i=1}^{n_{obs\ j}} [W_{obs\ lij}^2] - \left[\sum_{i=1}^{n_{obs\ j}} W_{obs\ lij} \right]^2 \right)^{-1} \\
 &\cdot \begin{pmatrix} \sum_{i=1}^{n_{obs\ j}} [W_{obs\ lij}^2] & - \sum_{i=1}^{n_{obs\ j}} W_{obs\ lij} \\ - \sum_{i=1}^{n_{obs\ j}} W_{obs\ lij} & n_{obs\ j} \end{pmatrix}
 \end{aligned}
 \tag{30}$$

The main diagonal elements in the second term of the product together with the determinant (first term of the product) in Eq. 30 govern the bias in the random intercepts and random slopes. We can simplify (30) further by noting that the determinant can be rewritten as (for readability we will drop the indices i, j , and 1 from here):

$$n_{obs} \cdot \sum (W_{obs}^2) - (\sum W_{obs})^2 = n_{obs}^2 \cdot var(W_{obs}),
 \tag{31}$$

where $var(W_{obs}) = 1/n_{obs} \cdot \sum (W_{obs} - \bar{W}_{obs})^2$. The proof is straightforward if we notice that this empirical variance can be rewritten as

$$\begin{aligned}
 var(W_{obs}) &= 1/n_{obs} \cdot \left\{ \sum (W_{obs}^2) - 1/n_{obs} \cdot (\sum W_{obs})^2 \right\} \\
 &= 1/n_{obs} \cdot \sum (W_{obs}^2) - 1/n_{obs}^2 \cdot (\sum W_{obs})^2 \\
 &\Rightarrow n_{obs}^2 \cdot var(W_{obs}) = n_{obs} \cdot \sum (W_{obs}^2) - (\sum W_{obs})^2
 \end{aligned}
 \tag{32}$$

After this simplification, it can be seen that the main diagonal elements of $(Z_j^{obs'} Z_j^{obs})^{-1}$ are $\sum (W_{obs}^2) / (n_{obs}^2 \cdot var[W_{obs}])$ and $(n_{obs} \cdot var[W_{obs}])^{-1}$.

We will start with the second component, the contribution to the random slopes bias. When the expected n_{obs} remains constant (as in our setting), the bias solely depends on $var(W_{obs})$. Note that as $|s|$ increases, it will be more likely that those Y -values with W_1 values in the tails of the distribution of W_1 will be deleted (see Fig. 1). For symmetric distributions such as the normal distribution that we use in our setting this implies that $var(W_{obs})$ will

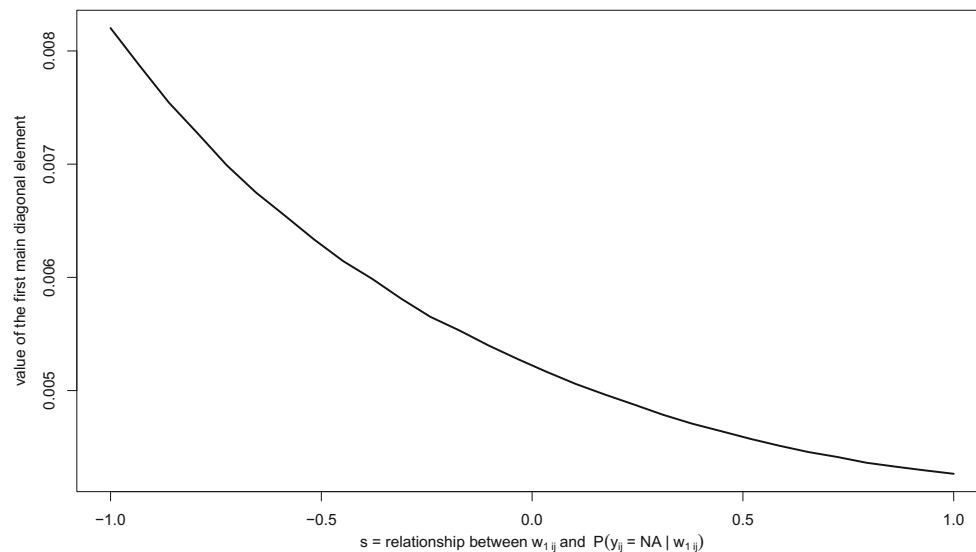


Fig. 6 Simulation of how $\sum (W_{obs}^2) / (n_{obs}^2 \cdot var[W_{obs}])$ changes as s takes different values

decrease and $var(W_{obs})^{-1}$ will increase and thus the bias in the variances of the random slopes will increase. This explains the U-shaped pattern seen in Fig. 3. Regarding the contribution to the random intercepts bias, we find that s has a negative relationship to the bias for our data setting. But whether $\sum (W_{obs}^2) / (n_{obs}^2 \cdot var[W_{obs}])$ is large or small highly depends on W_{obs} . We simulated this ratio for our data setup for various values of s . Figure 6 displays this ratio. The negative relationship between s and $\sum (W_{obs}^2) / (n_{obs}^2 \cdot var[W_{obs}])$ is in line with our empirical findings regarding the bias for the variance of the random intercepts (see Fig. 2). However, unlike the U-shaped relationship between s and the bias for the variance of the slopes that should hold for any random slopes model with a symmetric distribution of the slope variable, we emphasize that the relationship between s and the variance of the random intercepts is specific to this data setting and might well be reversed in other data settings.

References

- American Psychological Association (2017). Education and socioeconomic status. <http://www.apa.org/pi/ses/resources/publications/education.aspx>, accessed: 2017-06-22.
- Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53(1), 53–74. <https://doi.org/10.1002/bimj.201000140>.
- Asparouhov, T., & Muthén, B. (2010). Multiple imputation with mplus. MPlus Web Notes <https://www.statmodel.com/download/Imputations7.pdf>.
- Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4), 462–487. <https://doi.org/10.1177/0962280214521348>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., & Green, P. (2016). lme4: Linear Mixed-Effects Models using ‘Eigen’ and S4. <http://cran.r-project.org/web/packages/lme4/index.html>, r package version 1.1-12.
- Blossfeld, H. P., Roßbach, H.G., & Von Maurice, J. (2011). Education as a Lifelong Process - The German National Educational Panel Study (NEPS), vol. 14, Zeitschrift für Erziehungswissenschaft. <http://link.springer.com/journal/11618/14/2/suppl/page/1>.
- Brown, E., Graham, J., Hawkins, J., Arthur, M., Baldwin, M., Oesterle, S., Briney, J., Catalano, R., & Abbott, R. (2009). Design and analysis of the community youth development study longitudinal cohort sample. *Evaluation Review*, 33, 311–324.
- Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REAL-COM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45(5), 1–14. <https://doi.org/10.18637/jss.v045.i05>, <http://www.jstatsoft.org/v45/i05>.
- Clark NM, Shah S, Dodge JA, Thomas LJ, Andridge RR, & Little RJ (2010). An evaluation of asthma interventions for preteen students. *Journal of School Health*, 80(2), 80–87. <https://doi.org/10.1111/j.1746-1561.2009.00469.x>
- Diaz-Ordaz, K., Kenward, M. G., Gomes, M., & Grieve, R. (2016). Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Statistics in Medicine*, 35(20), 3482–3496. <https://doi.org/10.1002/sim.6935>.
- Drechsler, J. (2011). Multiple imputation in practice—a case study using a complex German establishment survey. *Advances in Statistical Analysis*, 95(1), 1–26.
- Drechsler, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1), 69–95. <https://doi.org/10.3102/107698614563393>.
- Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2), 222–240. <https://doi.org/10.1037/met0000063>.

- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
- Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Charles Griffin Book, Charles Griffin & Company, <https://books.google.de/books?id=UKzrAAAAMAAJ>.
- Goldstein, H. (2011). *Multilevel Statistical Models*, 4th edn., Wiley, Chichester, UK.
- Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1(4), 223–231. http://www.tandfonline.com/doi/abs/10.1207/S15328031US0104_02.
- Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behavior Research Methods*, pp 640–649, <https://doi.org/10.3758/s13428-015-0590-3>.
- Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer.
- Hedeker, D., & Gibbons, R. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psych Methods*, 2(1), 64–78. <https://doi.org/10.1037/1082-989X.2.1.64>.
- Horn, R. A., & Johnson, C. R. (1990). *Matrix Analysis*, reprint edn. Cambridge University Press, <http://amazon.com/o/ASIN/0521386322/>.
- Kenward, M. G., & Carpenter, J. (2007). Multiple Imputation: Current Perspectives. *Statistical Methods in Medical Research*, 16(3), 199–218. <https://doi.org/10.1177/0962280206075304>.
- Lenkeit, J. (2012). How effective are educational systems? A value-added approach to measure trends in pirls. *Journal for Educational Research Online*, 4(2), 143–173. <http://www.j-e-r-o.com/index.php/jero/article/view/317/157>.
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1), 141–165.
- McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton LS (2004a). *Evaluating Value-Added Models for Teacher Accountability*, 0th edn. RAND Corporation, <http://amazon.com/o/ASIN/0833035428/>.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67–101. <https://doi.org/10.3102/10769986029001067>.
- Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–573. <https://doi.org/10.1214/ss/1177010269>, <http://projecteuclid.org/euclid.ss/1177010269>.
- Mistler, S. A. (2013). A SAS Macro for Applying Multiple Imputation to Multilevel Data. *Proceedings of the SAS Global Forum* <https://support.sas.com/resources/papers/proceedings13/438-2013.pdf>.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects?. *Educational Evaluation and Policy Analysis*, 26(3), 237–257. <https://doi.org/10.3102/01623737026003237>, <http://epa.sagepub.com/content/26/3/237.short?rss=1&ssource=mfc>.
- O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data*. Information Age Publishing.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd edn. Cambridge University Press. http://prefixwww2.units.it/ipl/students_area/imm2/files/Numerical.Recipes.pdf.
- Quartagno, M., & Carpenter, J. (2016). jomo: A package for Multilevel Joint Modelling Multiple Imputation. <http://CRAN.R-project.org/package=jomo>, r package version 2.1-2.
- R. Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>.
- Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2), 143–149.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>, <http://biomet.oxfordjournals.org/content/63/3/581.short>.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9780470316696>.
- Schafer, J. L. (2016). pan: Multiple Imputation for Multivariate Panel or Clustered Data. <http://cran.r-project.org/web/packages/pan/index.html>, r package version 1.4.
- Scheffé, H. (1999). *The analysis of variance*, 1st edn. Wiley-Interscience.
- Siotani, M. (1967). Some applications of Loewner's ordering on symmetric matrices. *Annals of the Institute of Statistical Mathematics*, 19(2), 245–259. http://www.ism.ac.jp/editsec/aism/pdf/019_2_0245.pdf.
- Speidel, M., Drechsler, J., & Jolani, S. (2017). hmi: Hierarchical Multiple Imputation. [prefixhttps://CRAN.R-project.org/package=hmi](https://CRAN.R-project.org/package=hmi), r package version 0.7.4.
- StataCorp (2011). Accounting for clustering with mi impute. <http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/> (retrieved on 09.11.2016).
- Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, 50(3), 329–345. <https://doi.org/10.1002/bimj.200710423>.
- van Buuren, S. (2011). *Multiple imputation of multilevel data*, (pp. 173–196). Milton Park, UK: Routledge Academic. chapter 10.
- van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). mice: Multivariate imputation by chained equations. <https://cran.r-project.org/web/packages/mice/index.html>, r package version 2.25.
- Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New-York: Springer.
- Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*, 2nd edn. Cambridge, MA: The MIT Press.
- Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics*, 32(1), 231–256. <https://doi.org/10.1515/JOS-2016-0011>