Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 1, pp. 69–95
DOI: 10.3102/1076998614563393

Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity

Jörg Drechsler
Institute for Employment Research

Multiple imputation is widely accepted as the method of choice to address item-nonresponse in surveys. However, research on imputation strategies for the hierarchical structures that are typically found in the data in educational contexts is still limited. While a multilevel imputation model should be preferred from a theoretical point of view if the analysis model of interest is also a multilevel model, many practitioners prefer a fixed effects imputation model with dummies for the clusters since these models are easy to set up with standard imputation software. In this article, we theoretically and empirically evaluate the impacts of this simplified approach. We illustrate that the cluster effects that are often of central interest in educational research can be biased if a fixed effects imputation model is used. We show that the potential bias depends on three quantities: the amount of missingness, the intraclass correlation, and the cluster size. We argue that the bias for the random effects can be substantial while the bias for the fixed effects will be negligible in most real-data situations. We further illustrate this with an application using data from the German National Educational Panel Survey.

Keywords: *Imputation; fixed effects; missing data; multilevel; random effects*

1. Introduction

Most data sets used in education research show some form of natural clustering (students within classes, teachers within schools, students within universities, etc.) and there is general agreement among researchers that this clustering needs to be taken into account when analyzing the data. Generally, there are two approaches for this: including dummy variables for each cluster (fixed effects approach) or modeling the clusters as random (multilevel modeling/random effects approach). Researchers in education are often specifically interested in the relationship between the outcome and those variables that are constant within a cluster. For example, a researcher might be interested in how the attitude of the teacher affects the performance of the students. Since it is not possible to include variables in the model that are constant within a cluster with

the fixed effects approach (these variables would be perfectly collinear with the cluster dummies), multilevel models (Bryk & Raudenbush, 1992; Goldstein, 2011) are commonly applied in practice (see also Clarke, Crawford, Steele, & Vignoles, 2010, for a nice discussion of the pros and cons of fixed effects and multilevel models in different contexts).

However, despite the ubiquitous application of multilevel models in educational research, the implications of these models for nonresponse adjustments have never been thoroughly studied. Like any survey data, most data sets collected for educational research are affected by item nonresponse. In fact, the problem can be worse in hierarchically structured data since the nonresponse of a second level unit implies that the information of all the first level units is missing. For example, if a class does not participate in a test because the teacher refuses participation, no information on any of the students in that class can be collected for this test.

Multiple imputation (Rubin, 1978, 1987) is widely accepted as the preferable approach to deal with item nonresponse in surveys, but research on imputation strategies in the context of multilevel models is still limited. From a theoretical perspective, using a multilevel model at the imputation stage is recommended to ensure congeniality between the imputation model and the model used by the analyst (Meng, 1994). Uncongeniality can lead to biased results if the model used by the analyst is more complex than the imputation model and the imputation model omitted important relationships present in the original data. For this reason, different strategies to setup multilevel imputation models have been discussed in the literature (Liu, Taylor, & Belin, 1995; Schafer, 1997; Schafer & Yucel, 2002; Yucel, 2011).

However, these multilevel models are computationally more expensive since they require Markov Chain Monte Carlo approaches and only a limited number of software packages exist at present that allow researchers to utilize multilevel imputation models: The package pan in S-Plus by Joe Schafer and its replicate in R, also called pan, by Zhao and Schafer (2013), are based on the joint modeling approach. The joint modeling approach assumes that all variables follow a specific joint distribution, for example, a multivariate normal distribution. The multilevel modeling Software MLwiN also offers a multiple imputation macro based on this approach. Unlike the pan packages and the mice package described subsequently, MLwiN is capable of dealing with missing data on both levels of the model and also includes a multivariate probit model for imputing categorical variables (Carpenter, Goldstein, & Kenward, 2011). In practice, the assumption of a standard multivariate distribution is often too restrictive. Most surveys consist of a mix of categorical and continuous variables and skip patterns and logical constraints further complicate the modeling. Using a standard multivariate distribution in this context is often inappropriate (see Drechsler, 2011a, for further discussion of this issue). The multiple imputation package mice in R (van Buuren and Groothuis-Oudshoorn, 2011) is based on the more flexible sequential

regression multivariate imputation approach (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001) that defines a conditional model for each variable separately. This package also allows incorporating a hierarchical imputation model. However, imputations based on this model are limited to continuous variables, and the model allows a maximum of two levels. Furthermore, the sequential regression approach is strictly valid only if the conditional distributions are compatible, that is, if the iterative draws from the conditional distributions converge to draws from the joint distribution. This is not guaranteed in practice. However, Liu, Gelman, Hill, Su, and Kropko (2013) show that consistent results can still be obtained if the conditional models are correctly specified.

All other imputation procedures available in standard statistical software packages, such as SAS, Stata, and SPSS, assume that observations are independent and ignore intercluster correlations. As most applied researchers do not have the time, resources, or capabilities to implement their own imputation routines, a common approach for taking the hierarchy into account under this setup is to include indicator variables for clusters as fixed effects in the imputation model. This approach is also propagated on the frequently asked question website for the multiple imputation module in Stata (StataCorp, 2011). Since the fixed effects approach is easy to implement using standard imputation software, it has been widely used for the imputation of missing values in hierarchical data sets, (see, e.g., Brown et al., 2009; Clark et al., 2010; Si and Reiter, 2013).

The impact of using imputation models based on fixed effects when the analysis model is based on random effects has been studied in other contexts. Reiter, Raghunathan, and Kinney (2006) discuss how to incorporate the sampling design in multiple imputation. For clustered sampling designs, they compare three different imputation strategies: Ignoring the sampling design, using fixed effects, and using random effects to account for the sampling design. They investigate three scenarios. In the first scenario, both cluster effects and stratum effects are present in the data. In the second scenario, only stratum effects exist; and in the third scenario, the data show neither cluster effects nor stratum effects. In the simulation, the stratum effects are modeled as fixed whereas the cluster effects are modeled as random. They find that ignoring the sampling design for the imputation will always provide biased results except when neither cluster nor stratum effects are present. The fixed effects and the random effects imputation approaches both provide unbiased point estimates. However, the variances of the point estimates are overestimated with the fixed effects approach in Scenario 1 and especially in Scenario 3. Andridge (2011) addresses imputation approaches for cluster randomized trials. Cluster randomized trials are typically analyzed based on random effects models since according to the author "using fixed effects for clusters ... in *analysis* models leads to inflated type I error" (Andridge, 2011, p. 58). This article theoretically and empirically illustrates the bias of the variance estimates after multiple

imputation for the estimated fixed effects in random effects models when imputations are based on fixed effects. The data for the simulation are generated from a random intercept model. The parameters that vary between the simulations are the intraclass correlation, the cluster size and the missing data mechanism. The simulations show that the upward bias in the variance estimates decreases with increasing cluster size and increasing intraclass correlation. Whether the data are missing completely at random (MCAR) or missing at random (MAR) according to the definition of Rubin (1976) seems to have no impact on the bias. However, the theoretical results in Andridge (2011) are limited to the case without any covariates in the random effects model and even more important both articles only focus on the impact on the fixed effects and treat the random effects as nuisance parameters. This is justified in the applications considered by the authors (clustered sampling designs and cluster randomized trial), because in these applications the random effects are only included to get unbiased estimates for the fixed effects.

The situation is different in the educational context where the clustering effect is of direct interest. For example, the intraclass correlation, which is defined as the ratio of the unexplained variance on the cluster level (i.e., the variance of the random effects) divided by the total variance, measures to what extent unexplained differences in the dependent variable are due to cluster (i.e., school) effects. Correctly measuring the school effects is of great importance in education research. Thus, obtaining unbiased estimates for the random effects is more important in this research area than it is for the survey sampling and medical examples described previously. To our knowledge, the only article that addresses the impact of imputations based on fixed effects on inferences related to the random effects is by van Buuren (2011). However, that investigation is based only on simulations and no theoretical explanations for the observed biases are given. This article tries to fill this gap. In addition to deriving the theoretical bias that should be expected when the imputation models are based on fixed effects, we will discuss whether this bias is substantial enough to matter in practice. We focus only on random intercept models in this article. Extensions to random coefficient models should be straightforward and will be discussed briefly in the conclusions of the article.

The remainder of the article is organized as follows. The second section describes some general facts regarding the random intercepts model that will be important for understanding the consequences of the different imputation models that will be discussed in the third section. In the fourth section, a simulation study illustrates the impact of the fixed effects and random effects imputation methods on the parameters of the random intercept model. The fifth section evaluates the imputation impacts in practice based on data from the German National Educational Panel Study (NEPS). The article concludes with a general discussion of the findings, their implications, and possible areas for future research.

2. The Random Intercept Model

The random intercept model is a straightforward extension of the standard linear model. The major difference is that the intercepts for each cluster are allowed to vary randomly around the global intercept. The model is defined as:

$$Y_{ij} = \alpha_j + X_{ij}\beta + \varepsilon_{ij}, \quad \text{with} \quad \alpha_j \overset{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \psi^2), \quad \alpha_j \perp \varepsilon_{ij} \forall\ i,\ j,$$

where α_j is the random intercept in cluster j, $j = 1, \ldots, J$ and $X_{ij} = \{1, x_{1ij}, \ldots, x_{pij}\}$ is a $(1 \times (p+1))$ vector of explanatory variables for individual i, $i = 1, \ldots, n_j$, in cluster j, with n_j being the number of observations in cluster j; β is a $((p+1) \times 1)$ vector of regression coefficients and ε_{ij} is the error term. Finally, Y_{ij} is the outcome for individual i in cluster j and τ^2 and ψ^2 are the variances of the random intercepts and the error term, respectively. Under these model assumptions, the distribution of the random intercepts given the data and β, ψ^2, and τ^2 is given by (see, e.g., Gelman & Hill, 2007, p. 394):

$$\alpha_j | X, Y, \beta, \psi^2, \tau^2 \sim N\left(\mu_j, \eta_j^2\right), \text{ with}$$
$$\mu_j = \frac{\rho n_j \left(\bar{Y}_j - \bar{X}_j\beta\right) + (1-\rho)\cdot 0}{\rho n_j + (1-\rho)},$$
$$\eta_j^2 = \frac{\psi^2\tau^2}{n_j\tau^2 + \psi^2}$$

where $\rho = \tau^2/(\psi^2 + \tau^2)$ is the intraclass correlation.

We include $(1-\rho)\cdot 0$ in the expression for μ to illustrate that μ in fact is a composite estimator with relative weight $w_j = \rho n_j/\left(\rho n_j + (1-\rho)\right)$ given to the conditional cluster mean, $\bar{Y}_j - \bar{X}_j\beta$, and $1 - w_j$ given to zero. Looking at w_j, it is obvious that the expected shrinkage effect that pulls the random intercepts away from the conditional cluster mean toward zero depends on the cluster size n_j and the intraclass correlation ρ. Figure 1 illustrates the relationship between the different parameters. It depicts the relative weight given to the conditional cluster mean as a function of ρ and n_j. If the intraclass correlation is large, the weight increases quickly with the number of records in cluster j, that is, the shrinkage effect diminishes quickly. However, if ρ is small, the shrinkage toward zero can still be substantial even if the cluster size is large.

3. Imputation Strategies

The general procedures for obtaining multiply imputed data sets that correctly reflect the uncertainty from imputation can be summarized as follows: Let Z denote the set of variables for which information has been collected in a survey. With multiple imputation missing values in the collected data are replaced by repeated draws from $P(Z_{mis}|Z_{obs})$, where Z_{mis} denotes the missing part and Z_{obs} denotes the observed part of Z. Let θ be the set of parameters that govern the

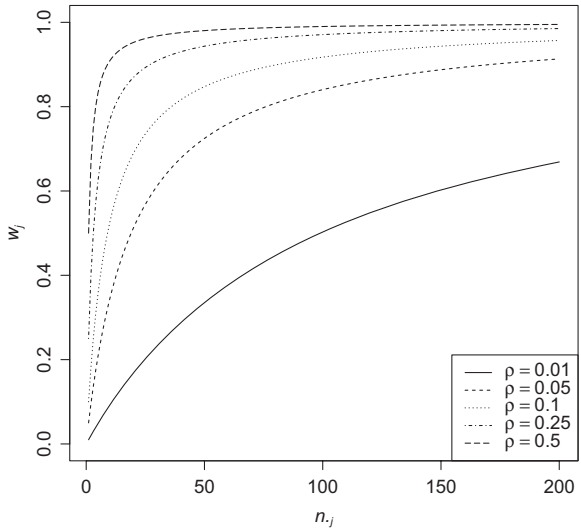

FIGURE 1. *Shrinkage effect for the random effect in a random intercept model as a function of the cluster size* n_j *and the intraclass correlation* ρ.

distribution of *Z*. Under the MAR assumption (Rubin, 1976), multiple imputations are generally generated in two steps:

1. Generate random draws for the set of parameters in θ from their observed-data posterior distributions $P(\theta|Z_{obs})$.
2. Generate random draws for Z_{imp} from its conditional predictive distribution $P(Z_{mis}|Z_{obs}, \theta)$ given the drawn parameters from Step 1.

As discussed in the introduction, two imputation strategies are commonly employed when dealing with missing values in hierarchical data sets: imputations based on fixed effects models including dummies for the clusters; and imputations based on multilevel models for which the clusters are specified as random effects. In the following subsections, we describe the models that are necessary for obtaining inferentially valid imputations for the two imputation approaches according to the two steps outlined previously.

3.1. The Fixed Effects Imputation Model

With the fixed effects imputation approach, the hierarchy in the data is reflected by incorporating dummy variables for each cluster. Let *X* be a $(n \times p)$ matrix of explanatory variables and let I_{ind} be an $(n \times J)$ indicator matrix identifying the cluster to which each record belongs with *n* being the number of

records and J being the number of clusters in the data. The assumed underlying model for the fixed effects imputation is given by:

$$Y = I_{ind}\alpha^f + X\beta^f + \varepsilon = R\gamma + \varepsilon, \quad \text{with } \varepsilon \sim N(0, I_n\sigma^2), \quad \alpha^f = \left(\alpha_1^f, \cdots, \alpha_J^f\right)',$$

where Y is the outcome variable, α^f is the $(J \times 1)$ dimensional vector of regression coefficients for the cluster specific intercepts, β^f is the $(p \times 1)$ dimensional vector of regression coefficients for the explanatory variables in X, $R = \{I_{ind}, X\}$, $\gamma' = \{\alpha'^f, \beta'^f\}$, and I_n is the $n \times n$ dimensional identity matrix. To keep the model identified, the global intercept needs to be dropped from the model since it would be perfectly collinear with the cluster indicator matrix.

Let $Y = \{Y_{obs}, Y_{mis}\}$, where Y_{obs} denotes the observed records in Y and Y_{mis} denotes the records for which Y is missing. Let $R = \{R_{obs}, R_{mis}\}$, $X = \{X_{obs}, X_{mis}\}$, and $I_{ind} = \{I_{ind,obs}, I_{ind,mis}\}$ be defined accordingly, that is, the subscript obs always defines the records for which Y is observed. Throughout this section, we assume that all records in X and R are fully observed or have been imputed in a previous step. Assuming noninformative priors for all parameters, imputations according to the fixed effects model defined previously can be generated based on the following two steps:

Step 1: obtaining values for $\theta = \{\sigma^2, \gamma\}$

- Draw $\sigma^2 | Y_{obs}, R_{obs} \sim (Y_{obs} - R_{obs}\hat{\gamma})'(Y_{obs} - R_{obs}\hat{\gamma})\chi^{-2}_{n_{obs}-p-J}$,
- Draw $\gamma | Y_{obs}, R_{obs}, \sigma^2 \sim N\left(\hat{\gamma}, \left(R'_{obs}R_{obs}\right)^{-1}\sigma^2\right)$,

where $\hat{\gamma}$ are the parameter estimates obtained from an ordinary least squares estimation, n_{obs} is the number of records for which Y is observed, and $\chi^{-2}_{n_{obs}-p-J}$ is a random draw from an inverse χ^2 distribution with $n_{obs} - p - J$ degrees of freedom.

Step 2: Obtaining values for Y_{mis}

- Draw $Y_{mis} | R, Y_{obs}, \theta \sim N(R_{mis}\gamma, \sigma^2)$.

3.2. The Random Effects Imputation Model

For the random effects imputation model, the assumed underlying model is given by:

$$Y = I_{ind}\alpha^r + X\beta^r + \varepsilon \quad \text{with} \quad \alpha^r \sim N(0, I_J\tau^2), \quad \varepsilon \sim N(0, I_n\psi^2), \quad \alpha^r = \left(\alpha_1^r, \cdots, \alpha_J^r\right),$$

Note that, for this model, the global intercept can still be included in X. Drawing the necessary parameters from their posterior distributions is more cumbersome for this model because the unconditional distributions for the parameters generally cannot be obtained in closed form. Thus, a Gibbs sampler is required

that iteratively draws from the conditional distributions of each parameter given all the other parameters. This Gibbs sampler needs to be run until convergence before any imputations can be obtained. Let $D_{obs} = \{Y_{obs}, X_{obs}\}$ denote the fraction of the data for which Y is observed. Assuming noninformative prior distributions, imputations under this model can be generated based on these two steps:

Step 1: Obtaining values for $\theta = \{\psi^2, \tau^2, \alpha_j^r, \beta^r\}$ by iteratively drawing from the following distributions until convergence:

- Draw $\alpha_j^r | D_{obs}, \psi^2, \tau^2, \beta^r, \sim N\left(\frac{n_{j,obs}\tau^2\left(\bar{Y}_{j,obs} - \bar{X}_{j,obs}\beta^r\right)}{n_{j,obs}\tau^2 + \psi^2}, \frac{\psi^2\tau^2}{n_{j,obs}\tau^2 + \psi^2}\right)$ for $j = 1, \ldots, J$,
- Draw $\beta^r | D_{obs}, \psi^2, \tau^2, \alpha^r \sim N\left(\left(X'_{obs}X_{obs}\right)^{-1}X'_{obs}\left(Y_{obs} - I_{ind,obs}\alpha^r\right), \psi^2\left(X'_{obs}X_{obs}\right)^{-1}\right)$,
- Draw $\psi^2 | D_{obs}, \tau^2, \alpha^r, \beta^r \sim \left(Y_{obs} - I_{ind,obs}\alpha^r - X_{obs}\beta^r\right)'\left(Y_{obs} - I_{ind,obs}\alpha^r - X_{obs}\beta^r\right)\chi^{-2}_{n_{obs}-1}$,
- Draw $\tau^2 | D_{obs}, \psi^2, \alpha^r, \beta^r \sim \left(\sum_{j=1}^{J}\left(\alpha_j^r\right)^2\right)\chi^{-2}_{J-1}$.

Step 2: Obtaining values for Y_{mis}

- Draw $Y_{mis} | X, \theta \sim N\left(I_{ind,mis}\alpha_j^r + X_{mis}\beta^r, \psi^2\right)$.

3.3. Implications of the Different Imputation Models

Comparing the two imputation strategies, it is evident why the first approach is usually preferred in practice: It is simpler to implement and no Markov Chain Monte Carlo methods are required. Since both models assume the same linear relationship between X and Y and the estimators for the fixed effects parameters are consistent for both models, we do not expect any biases for the estimates of β in the analysis model, no matter which imputation strategy is used. However, from the literature on random and fixed effects models, (see, e.g., Woolridge, 2010, chap. 10), we know that the random effects model is more efficient, that is, the estimates based on the random effects model have less variability. This fact is what causes the variance estimates to be positively biased if a multilevel model is used for analysis but missing values were imputed based on a fixed effects model. These findings are discussed in more detail in Reiter et al. (2006) and Andridge (2011), and we refer the interested reader to these articles. The focus of this article is on the impacts on the random effects estimates themselves, since these impacts have never been addressed before and are of great relevance in the educational research context.

It is generally difficult to compare the implications of the two different models directly since only the conditional distributions for all parameters are available in closed form for the multilevel imputation model. For this reason, we compare the conditional distribution of α for both models in this section. Under the fixed

effects imputation model, the conditional distribution of α_j^f given the data, β^f, and σ^2 is given by (see the Appendix for derivations):

$$\alpha_j^f | D_{obs}, \sigma^2, \beta^f \sim N\left(\bar{Y}_{j,obs} - \bar{X}_{j,obs}\beta^f, \sigma^2/n_{j,obs}\right),$$

Under the multilevel imputation model, the conditional distribution is given by:

$$\alpha_j^r | D_{obs}, \psi^2, \tau^2, \beta^r, \sim N\left(\frac{n_{j,obs}\tau^2\left(\bar{Y}_{j,obs} - \bar{X}_{j,obs}\beta^r\right)}{n_{j,obs}\tau^2 + \psi^2}; \frac{\psi^2\tau^2}{n_{j,obs}\tau^2 + \psi^2}\right).$$

The two models differ in two ways. First, the expected values of the cluster-specific intercepts are different. As discussed in the introduction, the α_j^r are pulled away from the cluster-specific intercepts α_j^f toward zero. Second, the variance of α_j is always larger in the first model. To see this, we note that $\sigma^2 = \psi^2$ which leads to (see the Appendix for derivations):

$$Var\left(\alpha_j^f | D_{obs}, \sigma^2, \beta^f\right) = \left(1 + \frac{1}{n_{j,obs}}\frac{(1-\rho)}{\rho}\right) Var\left(\alpha_j^r | D_{obs}, \psi^2, \tau^2, \beta^r\right). \quad (1)$$

As a consequence, the variances between the clusters and thus the intraclass correlation will be overestimated when missing values are imputed based on a fixed effects model. At the same time, the differences between the clusters as measured by the estimated α_js will also be overstated. This result is not surprising, as the expectation of the imputed values from the fixed effects model in cluster $j\left(Y_{j,imp}^f\right)$ given D_{obs} and β^f is as follows:

$$\begin{aligned} E\left(Y_{j,imp}^f | D_{obs}, \beta^f\right) &= E\left(\iota_{j,mis}\alpha_j^f + X_{j,mis}\beta^f + \varepsilon_{j,mis}\right) \\ &= \iota_{j,mis}\left(\bar{Y}_{j,obs} - \bar{X}_{j,obs}\beta^f\right) + X_{j,mis}\beta^f, \end{aligned}$$

where $\iota_{j,mis}$ is a vector of ones with length equal to the number of missing records in cluster j. The same expectation for the imputed values from the random effects model $\left(Y_{j,imp}^r\right)$ is given by:

$$\begin{aligned} E\left(Y_{j,imp}^r | D_{obs}, \beta^r\right) &= E\left(\iota_{j,mis}\alpha_j^r + X_{j,mis}\beta^r + \varepsilon_{j,mis}\right) \\ &= \iota_{j,mis}\frac{n_{j,obs}\tau^2\left(\bar{Y}_{j,obs} - \bar{X}_{j,obs}\beta^r\right)}{n_{j,obs}\tau^2 + \psi^2} + X_{j,mis}\beta^r. \end{aligned}$$

Since the estimator for the fixed effects is consistent in both models and thus $E(\beta^f) = E(\beta^r)$ this leads to

$$E\left(Y_{j,imp}^r | D_{obs}, \beta^f\right) = w_{j,obs}E\left(Y_{j,imp}^f | D_{obs}, \beta^r\right).$$

This result implies that the imputed values based on the multilevel imputation model are in expectation always closer to zero than the imputed values based on the fixed effects imputation. Or, expressed differently, within each cluster the imputed values from the fixed effects imputation model will vary randomly around the conditional cluster mean whereas the imputed values from the multilevel model will vary randomly around the shrunken cluster mean. The shrinkage effect pulls the conditional cluster means closer to zero for all clusters in which some missing values were imputed. This is what we would expect under the multilevel model assumed by the analyst. The multilevel model implies that the cluster effects have an expectation of zero and thus we expect the estimated cluster effects to be closer to zero with increasing sample size. This shrinkage effect is lacking in the fixed effects imputation which explains the biases discussed above. Again, we note that the difference between the two models decreases with increasing ρ and with the number of observed records in the cluster.

4. Simulation Study to Evaluate the Practical Implications

From the previous sections, it is obvious that using a fixed effects imputation model induces biases for various components of the multilevel analysis model. The important question that still needs to be addressed is whether the amount of bias is large enough in practice to justify the computational burdens of the random effects imputation model. The formulas derived previously only illustrate the differences of fitting a fixed effects model or random effects model to the data. In the missing data context, the impact of the chosen model also depends on the amount of missing data. The impact of the missingness rate is 2-fold. Clearly, the impact of any imputation model increases with the amount of missing data. If only a small fraction of the data is missing, the impacts on the final results will be minor no matter which imputation model is chosen. But in the multilevel context, the missing rate has another more direct impact on the results. From Equation 1, we know that the difference of the variances depends on the number of observed records in each cluster. Thus, if the missingness rate in a cluster is large, the two models will differ more. Since it is difficult to quantify the impact on the analysis results analytically, we conduct a simulation study in which we vary the three parameters that directly influence the results: the intraclass correlation ρ, the number of records per cluster n_j, and the missingness rate.

4.1. Simulation Design

For the simulation, we assume that the analysis model of interest is a random intercept model given by:

$$Y_{ij} = \alpha_j + \beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + \varepsilon_{ij}, \quad \text{with} \quad \alpha_j \overset{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \psi^2), \quad \alpha_j \perp \varepsilon_{ij} \forall\ i,\ j,$$

where X_{ij} and Y_{ij} vary on the individual level, whereas Z_j only varies between clusters. We further assume that the data are generated based on the following distributions:

$$\begin{aligned} X_{ij} &\sim N(0,1), \\ Z_j &\sim N(0.5,1), \\ \alpha_j &\sim N(0,2\rho), \\ Y_{ij}|\alpha_j, X_{ij}, Z_j &\sim N\big(\alpha_j + 0.5 - 0.5X_{ij} + 2Z_j, 2(1-\rho)\big). \end{aligned}$$

We repeatedly generate samples from these distributions, induce missingness, and impute the missing values according to the fixed effects imputation model described previously (note that we either need to drop the global intercept and Z_j or we need to drop two of the cluster-specific intercepts in the fixed effects imputation model to keep the model identified). We also imputed the missing values based on a multilevel model. As expected, all parameter estimates from the imputed data were unbiased in this case and we omit all results for brevity. The parameters that vary between the different simulation runs are as follows:

- The intraclass correlation: $\rho = \{.05, 0.1, \ldots, 0.5\}$.
- The number of records in each cluster: $n_j = \{10, 15, 25, 50\}$.
- The probability for Y to be missing: $P(Y = \text{missing}) = \{0.05, 0.1, 0.25\}$.

We fix the number of clusters to be 25 in all simulation runs and repeat the whole process of sampling, inducing missingness and imputing the missing values 1,000 times for each parameter combination. For simplicity, we use a missing data mechanism that implies that the probability for Y to be missing is constant across all records, that is, the mechanism is based on the MCAR assumption (Rubin, 1976). This assumption is often questionable in practice. However, as the results in Section 3.3 show, the bias in the random effects inferences does not depend on the data except through $n_{j,obs}$. Thus, the results would not change if the missingness would depend on X. This fact was also confirmed in a small simulation study not shown here for brevity, and these findings are consistent with the findings in Andridge (2011). In her simulations, the author found no difference in the bias for the fixed effects if she induced missings based on the MAR assumption instead of the MCAR assumption. The implications would be different, however, if the missingness would depend on the cluster variable Z. Obviously, if the missingness mechanism is such that the probability for Y to be missing is higher for cluster a than for cluster b, this would imply that if a fixed effects imputation model is used and the two clusters are of equal size, the estimated random effect of cluster a and its conditional variance would be more biased since $n_{j,obs}$ would be smaller. Nevertheless, the implications for global measures such as the intraclass correlation are not as obvious. Evaluating the impacts of a missingness mechanism that depends on the cluster-level variables

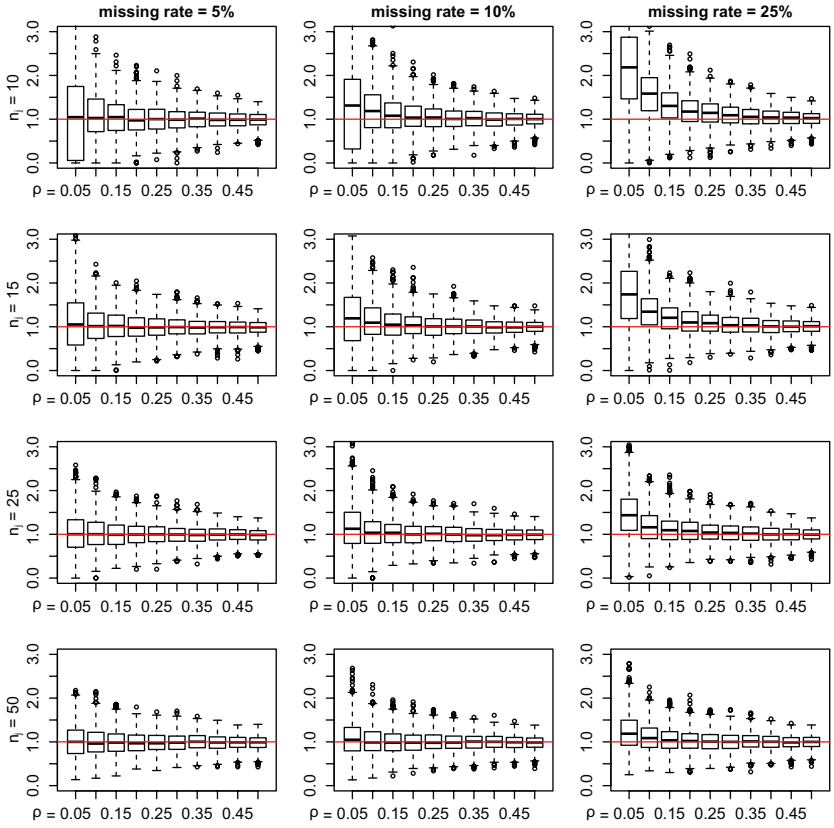

FIGURE 2. *Boxplots of the estimated intraclass correlation* $\hat{\rho}$ *from 1,000 simulation runs divided by the true intraclass correlation* ρ. *For unbiased estimates, the boxplots should be centered at 1 as indicated by the solid line in each panel.*

is beyond the scope of this article and would be an interesting area for future research.

The parameters of interest in the simulation are the estimated intraclass correlation and the estimated variances of the fixed effects $\beta_1, \ldots, \beta_3$ since we expect these quantities to be biased. The results for the intraclass correlation are reported in Figure 2. The figure depicts boxplots of the estimated intraclass correlation $\hat{\rho}$ divided by the true intraclass correlation from the data-generating process. All boxplots are based on 1,000 simulation runs. If the estimate of the intraclass correlation is unbiased, the boxplots should be centered on one as indicated by the black line in each panel. We limit the range of the displayed ratio to a maximum of 3 for all panels to allow for an easy comparison of the results across

panels. As a consequence, some extreme simulation results are outside the reported range in the figure.

The results are consistent with the analytical results described in Section 3.3. The intraclass correlation tends to be overestimated if a fixed effects imputation model is used. This is especially evident for the last column of panels for which the missing rate is set to 25%. For this column, the other two influencing factors are also easy to identify. In each panel, the extent of overestimation decreases as the true ρ increases. At the same time, an increasing number of observations in each cluster also reduces the amount of overestimation. However, the simulations also illustrate that the negative effect of the misspecified imputation model strongly depends on the amount of missing data. If the missingness rate is 10%, a positive bias can only be observed if $\rho \leq .15$ for $n_j \leq 15$ and $\rho \leq .05$ for $n_j \leq 25$. If the missingness rate is 5%, no biases are observable for any ρ/n_j-combination.

Figure 3 contains the results for the estimated variance of $\hat{\beta}_3$ (the results for $\hat{\beta}_1$ and $\hat{\beta}_2$ are similar and are thus omitted for brevity). The boxplots contain the ratios of the estimated variance of the parameter divided by the true variance. The true variance is computed as the empirical variance of $\hat{\beta}_3$ across the 1,000 simulation runs. The plotted ratios are limited to a maximum of two to allow for an easy comparison of the results across panels.

The general findings are similar to the findings for the intraclass correlation. The variance tends to be overestimated, and the overestimation decreases with an increasing intraclass correlation and/or cluster size. A decreasing missingness rate also reduces the positive bias. But the bias is generally less severe than the bias for the intraclass correlation. We don't observe any bias with a missingness rate $\leq$10% and, even for a missingness rate of 25%, the bias is only observable for small ρ/n_j-combinations.

4.2. Discussion

The empirical results indicate that the general advice to simply use a standard imputation model with cluster indicators included as dummies in the model (fixed effects imputation) often seems to be a feasible solution in practice. For most real data sets, the missingness rate per variable seldom exceeds 10% and intraclass correlations in the educational context are often found to be between 0.05 and 0.2 (Snijders & Bosker, 1999, p. 46). In this case, the bias induced by the fixed effects model is so small that it seems justified to use the fixed effects imputation model for convenience. Furthermore, if the clustering is at the school level, many surveys will contain 50 observations or more for each school and in this case the fixed effects imputation will always provide nearly unbiased results irrespective of the missingness rate or the level of intraclass correlation.

However, the random effects imputation model might still be preferable in some settings for several reasons: First, the random effects imputation model

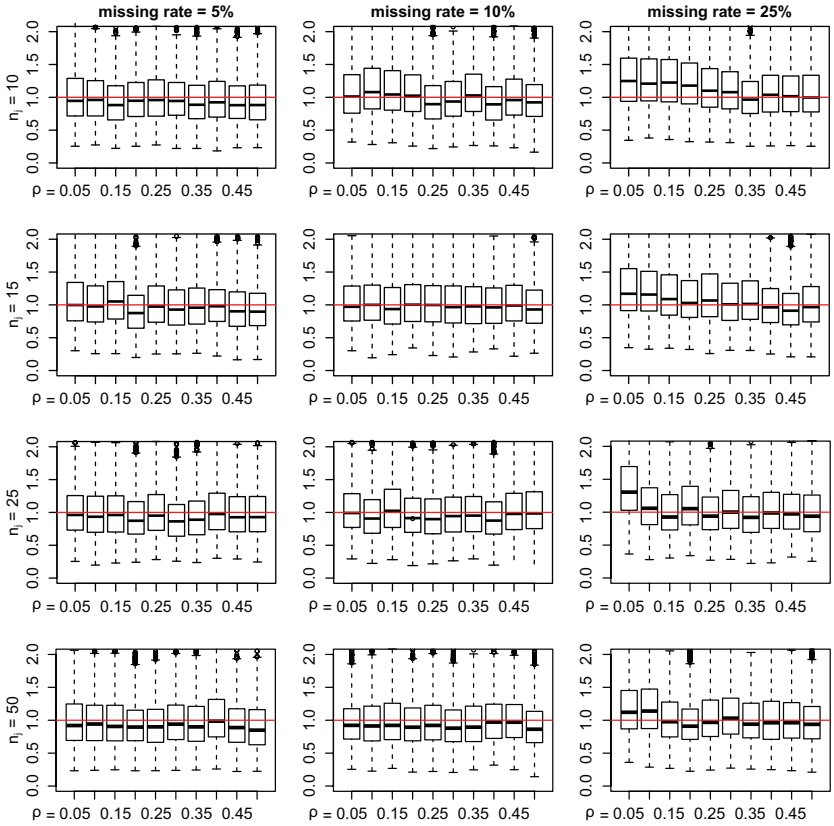

FIGURE 3. *Boxplots of the estimated variance of* $\hat{\beta}_3$ *divided by the empirical variance of* $\hat{\beta}_3$ *across the 1,000 simulation runs. For unbiased estimates, the boxplots should be centered at 1 as indicated by the solid line in each panel.*

is guaranteed to produce unbiased results in all settings (unless the MAR assumption is violated or the analysis model is misspecified, in which case any discussion about the differences between the two imputation models is moot since both models will lead to biased results). Second, the general advice in multiple imputation is to include as many variables as possible in the imputation model to avoid biased estimates obtained from the imputed data (Little & Raghunathan, 1997). This strategy is limited with the fixed effects model since (a) the inclusion of dummies for all clusters already requires $m - 1$ parameters to be estimated (adding more variables might cause multicollinearity problems) and (b) no variables that are constant within clusters can be included in the imputation model beyond the cluster specific dummies. Third, in some contexts, the "missingness rate" could be large. For example, in some studies, some variables are only gathered

for a small subset of the original sample if the information is expensive to obtain. This could be the case if expensive tests need to be conducted for those variables. It sometimes makes sense to impute the missing information for all those units that were not selected for the additional measurements (see, e.g., Schenker, Raghunathan, and Bondarenko, 2010). Obviously, the missingness rates are much larger in this case. Another example is the release of synthetic data for confidentiality reasons (Drechsler, 2011c). The idea of synthetic data is closely related to the idea of multiple imputation for missing data. The only difference is that sensitive values instead of missing values are replaced by multiple draws from predictive distributions given the observed data. In most synthetic data applications, the values for all records are replaced by synthetic values at least for some of the variables (Abowd, Stinson, & Benedetto, 2006; Drechsler, 2011b; Drechsler, Bender, & Rässler, 2008; Drechsler, Dundler, Bender, Rässler, & Zwick, 2008; Drechsler and Reiter, 2010, 2011; Kinney et al., 2011). This means that 100% of the records are imputed and using a fixed effects imputation model would cause substantial bias in this case.

5. Multiple Imputation in the NEPS

We evaluate the impact of the two different imputation models using data from the NEPS. 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 obtain 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, and so on. 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. See Blossfeld, Roßbach, and von Maurice (2011), for further details.

For our evaluation, we use data from the fourth cohort: students in upper secondary school. The starting cohort collected in 2010 contains about 14,500 ninth graders from 500 different schools. The data were collected in a two-stage sampling design. First, a stratified sample of schools was drawn from a register of all schools in Germany. In the second stage, two classes were sampled from each school (see Aßmann, Steinhauer, & Zinn, 2012, for further details on the sampling design). Questionnaires were given to the students, their parents, their German and Mathematics teachers, and the principals of the school. Furthermore, the students participated in competence tests. The collected data provide a rich source of information from demographic variables such as migration background or level of education of the parents over the learning environment at home, the occupational plans and educational aspirations to motivational aspects and personality traits. The teachers and principals provide further information such as

the social class and the migration background at the class and school levels. See von Maurice, Sixt, and Blossfeld (2011), for further details on the survey of the fourth cohort of the NEPS.

Four waves of the survey of the fourth cohort are available to the scientific community so far. External researchers have three different options for how to access the data: using scientific use files disseminated by the Leibniz Institute for Educational Trajectories, remote access or on-site access. The amount of information that is available to the researcher increases in the same order. See the website of the NEPS research data center (https://www.neps-data.de/en-us/datacenter/dataaccess.aspx) for further information on how to access the data. We use the data on the competence tests and the student questionnaires available in the scientific use file. We examine whether the self-evaluation of the students regarding their own competencies and interests in mathematics has a significant influence on their performance in the mathematics competence test beyond their grade in mathematics in the report card from the previous year. The model is specified as:

$$lit_m = I_{ind}\alpha_j + \beta_0 + \beta_1 \cdot self_m + \beta_2 \cdot help_m + \beta_3 \cdot interest_m + \beta_4 \cdot grade_m + \gamma \cdot Z + \varepsilon,$$
$$\alpha_j \sim N(0, I_J\tau^2), \qquad \varepsilon \sim N(0, I_n\psi^2),$$

where lit_m is the mathematical competence measured as mathematical literacy and α_j is a random intercept that varies for each class. Measures of the mathematical literacy are provided in form of weighted maximum likelihood estimates (WLEs; see Pohl & Carstensen, 2012, and Duchhardt & Gerdes, 2013, for details). The variables $self_m$, $help_m$, and $interest_m$ are each based on the average score on several Likert-scaled questions for which the students should comment on how much different statements apply to them. The variable $self_m$ is a measure for the self-evaluation regarding mathematics. The underlying questions ask whether the student gets good grades in mathematics, whether mathematics is one of their best subjects, and whether the student feels that he or she has always been good at mathematics. The variable $help_m$ measures the helplessness that students feel toward mathematics. Topics include grades, resignation regarding class tests, unfulfilled expectations, being asked to answer questions, and homework. The variable $interest_m$ is a measure for the subject-related interest in mathematics and is based on questions regarding considering mathematics as fun, feeling that time flies when working on mathematical problems, willingness to spend leisure time on mathematics, and whether mathematics is considered important. The variable $grade_m$ is the grade in mathematics in the report card from the previous year. Finally, Z contains a set of control variables, namely, *sex*, *type of school*, and *year of birth*. It should be noted at this point that the model only serves as an illustration for the impacts of the different imputation models. The model could certainly be improved by including more control variables and information from the questionnaire of the parents and teachers to make the

TABLE 1
Variables Included in the Final Imputation Model

Variable	Description	Characteristics	Missing(%)
Lit_m	Mathematical literacy	Range: −4.37–4.62	3.84
$self_m$	Mathematical self-evaluation	Range: 1–4	6.61
$help_m$	Mathematical helplessness	Range: 1–4	11.46
$interest_m$	Interest in mathematics	Range: 1–4	11.61
$grade_m$	Grade from latest report card	6 categories (1 = best grade)	8.35
sex		binary	0
type of school		5 categories	0
year of birth		3 categories (<1995, 1995, >1995)	0
classID	Unique ID for each class	1,004 categories	0

assumptions of the random effects model more plausible. Furthermore, plausible values instead of the WLEs should be used for the competence measures in practice to account for measurement error. We also ignore the complex sampling design of the survey. However, including more variables, working with multiple plausible values, and accounting for the sampling design would only complicate the imputation task without adding any additional insights regarding the imputation model implications.

We use the first wave of the data and only include students who filled out the questionnaire in the first wave. This reduces the initial sample size of 16,425 records to 16,254. We also exclude information from schools for children with special needs since the survey design for these schools is different and it is generally recommended to analyze the data from these schools separately from the rest of the data. This step further reduces the sample size to 15,099. Finally, we also drop all classes for which any of the variables has no observed values since these missing values cannot be imputed with the fixed effects imputation approach. Overall eight classes containing 29 students need to be dropped leading to a final sample size of 15,070 records. Note that these missing records could still be imputed using a multilevel imputation model (this is another disadvantage of the fixed effects imputation). We only drop those records for both models to make the results comparable. The average cluster size in the final sample is 15.01. Table 1 provides a basic description of the variables used.

In the analysis model, all math-related variables are treated as continuous except for $grade_m$. This variable and all control variables are treated as categorical. The *classID* is used to identify the random intercepts. The model is fit using the lmer function in *R*.

The missing values in the math-related variables are imputed by sequential regression, that is, for each variable, a conditional model given all the other

variables are estimated and imputed values are drawn from this model. For simplicity, all variables including the $grade_m$ variable are treated as continuous in the imputation models. For $grade_m$, the imputed values are rounded to the closest feasible integer after imputation. A similar approach has been suggested by Schafer and Olsen (1998), but see Yucel, He, and Zaslavsky (2008) for potential problems with this imputation strategy. We recommend using an ordered probit model in practice, but we don't expect any negative impacts of this simplification for our comparison since both imputation models are affected similarly. We generate $m = 15$ imputed data sets. For the multilevel model, we run three independent chains and store five imputations from each chain. We run the Gibbs sampler for 20 iterations between each imputation and set the burn in phase of the Gibbs sampler to 99 iterations. We store the mean and the variance of the imputed values for each variable at each iteration. Several convergence diagnostics, such as traceplots, autocorrelation functions, and the Gelman and Rubin convergence diagnostic (Gelman & Rubin, 1992) based on these estimands indicate that 99 iterations are sufficient to ensure convergence and 20 iterations between the draws guarantee independence of the draws.

Results for the model of interest based on the two different imputation strategies are reported in Table 2. The table reports the point estimates, their standard errors (in brackets), and the ratio of the 95% CI lengths with the lengths of the estimates based on the multilevel imputation model in the denominator. The last column of the table measures to what extent the CIs of the estimates from the two models overlap. The CI overlap measure used here was originally suggested by Karr, Kohnen, Oganian, Reiter, and Sanil (2006) to evaluate the analytical validity of data sets that have been protected by some statistical disclosure limitation technique such as swapping. In the data confidentiality context, the measure is computed as follows: For any estimand, compute the 95% CIs for the estimand from the protected data, (L_p, U_p), and from the original data, (L_o, U_o). Then, compute the intersection of these two intervals, that is, compute the endpoints of the overlap between the two CIs (L_i, U_i). The utility measure is as follows:

$$IO = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_p - L_p)}. \tag{2}$$

When the intervals are nearly identical, corresponding to high analytical validity, $IO \approx 1$. When the intervals do not overlap, corresponding to low analytical validity, $IO = 0$. The second term in Equation 2 is included to differentiate between intervals with $(U_i - L_i)/(U_o - L_o) = 1$ but different lengths. For example, for two protected data intervals that fully contain the collected data interval, the measure IO favors the shorter interval. The protection method is successful in terms of preserving analytical validity if large values of IO are obtained for many estimands.

TABLE 2
Comparison of Parameter Estimates and Their Standard Errors (in Parentheses) for a Random Intercept Model (Dependent Variable: Mathematical Literacy) Fitted on Two Imputed Data Sets

Variable	Fixed Effects	Multilevel	CI Ratio	*IO*
$self_m$	0.210	0.211	1.090	.959
	(1.294×10^{-3})	(1.187×10^{-3})		
$help_m$	−0.062	−0.065	1.062	.946
	(1.261×10^{-3})	(1.187×10^{-3})		
$interest_m$	0.023	0.019	0.994	.914
	(1.122)	(1.129)		
$grade_m = 2$	−0.474	−0.477	1.005	.979
	(3.111×10^{-3})	(3.095×10^{-3})		
$grade_m = 3$	−0.735	−0.741	1.015	.952
	(3.337×10^{-3})	(3.289×10^{-3})		
$grade_m = 4$	−0.838	−0.841	1.040	.979
	(3.843×10^{-3})	(3.696×10^{-3})		
$grade_m = 5$	−0.838	−0.838	1.011	.995
	(4.583×10^{-3})	(4.535×10^{-3})		
$grade_m = 6$	−0.808	−0.782	0.968	.951
	(13.157×10^{-3})	(13.590×10^{-3})		
ρ	.198	.192	1.030	.000
	(0.156×10^{-3})	(0.147×10^{-3})		

Note. Cohort 4 of the NEPS; doi:10.5157/NEPS:SC4:1.1.0. Missing values are imputed based on a fixed effects model or a multilevel model. The third column reports the ratio of the 95% confidence intervals (CI). The last column reports the overlap of the 95% CI obtained from the two models.

The context is different here, but if we replace *protected data* with *imputed data based on fixed effects*, *original data* with *imputed data based on multilevel models*, and *analytical validity* with *similarity*, the measure can easily be transferred to our setting. We know that the multilevel imputation model is the "correct" model because it is congenial to the analysis model. Thus, we can interpret *IO* as a measure that helps to quantify the negative impact of using a fixed effects model since it measures to what extent we would obtain similar results to the "true" model (again, we ignore that the multilevel model might be misspecified, too, because a comparison of the models would not provide any useful insights if we assume that the analysis model is misspecified).

For both models, the results are in line with what we would expect. There is a strong negative relationship between the performance in the competence test and the grades (the fact that the negative effect for $Grade_m = 6$ is estimated to be smaller than for $Grade_m = 5$ is probably due to a small sample size for $Grade_m = 6$). Only 37 students reported this grade. It is interesting to see that all self-evaluation measures have a significant effect on the performance beyond the

grade. Most effects are rather small, but the estimated effect of $self_m$ is strong enough that it could possibly compensate any negative effect of the grade (note that the range of $self_m$ is from one to four, which means that the predicted performance score for a very self-confident student can increase by up to 0.84).

Of course, in our context, we are more interested in the impact the different imputation models have on the results. For the fixed effects, we hardly find any differences between the two imputation models. We expect the point estimates to be similar since the choice of the imputation model should only affect the estimated variances. But the estimated standard errors of the point estimates are also very close even though the missingness rate is more than 10% for some of the variables. All CI ratios and CI overlap measures are close to one, indicating that similar inferences would be obtained for all estimated fixed effects. However, this is not true for the random effects–based inference. The estimated intraclass correlation differs between the two imputation methods. As expected, the estimated intraclass correlation is larger (.198 compared to .192) if the fixed effects imputation is used. The standard errors based on the fixed effects imputation model are only slightly larger than the standard errors from the multilevel model leading to a CI ratio of 1.03. Arguably, the difference in the point estimates for the intraclass correlation is small (.006) in this application. Thus, from an applied perspective, the fixed effects results might be acceptable since the drawn conclusions would hardly differ based on these results. Still, the CI overlap for the intraclass correlation is zero indicating considerable difference in the inference obtained for this estimand depending on which imputation model is used. Thus, if interest lies in the random effects, a multilevel imputation model should be preferred, to avoid biases in the obtained inferences. We note that the distribution of the estimated intraclass correlation is not normal and thus it is not valid to use the multiple imputation combining rules to estimate the variance for this estimand (Zhou & Reiter, 2010). As a proxy, we use the estimated intraclass correlation computed at each iteration of the sequential regression imputation after the burn-in phase. Since we run each chain for 81 iterations after burn-in to acquire five imputations, we obtain 243 estimates for the intraclass correlation. We treat these estimates as a proxy for the posterior distribution of the intraclass correlations and use the empirical standard deviation across these estimates as an estimate for the true standard deviation. Similarly, we use the empirical 2.5% and 97.5% quantiles of the estimates when calculating the CI ratio and the overlap measure.

The findings for this application are in line with the results from the simulation study. We did not find any bias for the variance of the fixed estimates with a missing rate of approximately 10%, an average cluster size of 15 and an intraclass correlation of approximately .2. But we noticed a small bias for the intraclass correlation based on these parameter constellations in our simulations.

6. Conclusion

Multilevel models play a vital role in educational research. To address the item nonresponse in the data, a common recommendation, which is often adopted by practitioners because of its simplicity, is to use standard linear regression models with dummies to identify the clusters. While this approach is clearly inferior to using multilevel models at the imputation stage, the theoretical and practical implications of the simplified approach have never been thoroughly studied. In this article, we illustrated that the approach indeed causes some biases not only for the estimated variance of the fixed effects estimates but also for the random effects which are especially important in the educational context. Both the estimated cluster effects as well as the variance between the clusters will be too large, that is, the influence of the cluster (the class or the school effect) will generally be overestimated. We also evaluated through extensive simulations and real-data applications how relevant the bias is in practice. We found that unless the missing rate is very large and/or the intraclass correlation is very low and the number of records in the cluster is small, the bias for the variance of the fixed effects is small. Thus, if researchers are only interested in inferences on the fixed effects, the fixed effects imputation approach can be an easily implemented alternative. The bias for the random effects is more substantial. In our application based on the NEPS data, the estimated intraclass correlation changed from .198 to .192. While this difference is still small in absolute terms the CIs for the intraclass correlation showed no overlap indicating substantial differences in the inferences drawn from the two imputed data sets.

Beyond the potential bias, there are other reasons why a multilevel imputation model should be preferred if possible. First, the model is guaranteed to produce unbiased results even in extreme cases with very high missingness rates and/or small intraclass correlations. Second, with the multilevel model, it is possible to include additional variables that are constant on the cluster level, for example, information on the teachers if the cluster levels are chosen to be students within classes or background information on the schools if the clustering is modeled as students within schools. This approach is not possible with fixed effects imputation. While this limitation should not be problematic in terms of bias (the combined school effects are still modeled through the cluster specific intercepts), using this additional information will make the random effects imputation model more efficient.

Nevertheless, researchers might not be able to implement a multilevel imputation model for various reasons: Multilevel models are computationally intense, and applying these models to all variables with missing data might be too complex. Furthermore, software for multilevel model imputation is still sparse and often relies on questionable model assumptions and applied researchers might not have the time or capability to implement their own imputation models.

Because information on the parameters—n_j, ρ, and the missingness rate—that drive the bias of the fixed effects imputation approach are available from the

observed data, researchers can judge based on the findings in this article whether the bias is expected to be relevant. For inferences on the fixed effects, we expect that this will hardly ever be the case in most practical applications, since the intraclass correlation will seldom be below .1, cluster sizes are usually larger than 10, and the missing rate is seldom above 10% to 15%. Under these circumstances, researchers can use the more convenient fixed effects imputation approach and need not be concerned about potential biases. However, if random effects–based inferences such as the intraclass correlation are important, researchers should consider carefully whether a fixed effects imputation model is appropriate.

We only focused on random intercept models in this article. However, we believe that similar arguments could be made for random coefficient models in general. Since the shrinkage effect is ignored during the imputation, the variance of the random intercepts and the random slopes will generally be overestimated. Whether the bias could also be ignored in practice is a question for future research. However, it should be noted that implementing a fixed effects imputation model is also more complicated in this case. Beyond the cluster indicators, the model also needs to include an interaction term between the cluster variable and any variable that is allowed to vary randomly in the random coefficient model. Fitting models with these many parameters can be challenging in practice. Furthermore, as one of the referees pointed out, this approach can only be applied directly if the variables that should be interacted with the cluster variable are fully observed. Otherwise, multiple sets of product terms need to be included further complicating the fitting of the models.

Appendix

1. Derivation of $\alpha_j^f | \boldsymbol{D_{obs}}, \sigma^2, \beta^f$

Let $\gamma' = (\alpha'^f, \beta'^f)$ and $R = (I_{ind}, X)$ with dimension $n \times (J + p)$, where n is the number of observations in the data, J is the number of clusters, and p is the number of covariates in X. From standard linear regression results, we know that using uninformative priors $\gamma | D_{obs}, \sigma^2 \sim N(\hat{\gamma}, \sigma^2 (R'R)^{-1})$ which is equivalent to

$$\begin{pmatrix} \alpha^f | D_{obs}, \sigma^2 \\ \beta^f | D_{obs}, \sigma^2 \end{pmatrix} \sim N(\mu, \Sigma), \quad \text{with}$$

$$\mu = \begin{pmatrix} \hat{\alpha}^f \\ \hat{\beta}^f \end{pmatrix}$$

$$\Sigma = \sigma^2 \begin{pmatrix} (R'R)^{-1}_{(1:J)\times(1:J)} & (R'R)^{-1}_{(1:J)\times(J+1:J+p)} \\ (R'R)^{-1}_{(J+1:J+p)\times(1:J)} & (R'R)^{-1}_{(J+1:J+p)\times(J+1:J+p)} \end{pmatrix}$$

$$= \sigma^2 \begin{pmatrix} (R'R)^{-1}_{11} & (R'R)^{-1}_{12} \\ (R'R)^{-1}_{21} & (R'R)^{-1}_{22} \end{pmatrix}.$$

For multivariate normal distributions, it generally holds that if

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

then

$$X_1|X_2 \sim N\big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big).$$

Thus, the conditional distribution for α^f is given by:

$$\alpha^f|D_{obs}, \sigma^2, \beta^f \sim N\left(\hat{\alpha}^f, (R'R)_{11}^{-1} - (R'R)_{12}^{-1}(R'R)_{22}(R'R)_{21}^{-1}\right). \tag{3}$$

It also holds (see, e.g., Harville, 1999, p. 100) for any arbitrarily partitioned nonsingular matrix $M = \begin{pmatrix} T & U \\ V & W \end{pmatrix}$, for which the inverse matrix $B = \begin{pmatrix} T & U \\ V & W \end{pmatrix}^{-1}$ is partitioned as $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ and each partition has the same dimensions as the partitions in M that

$$B_{11} = T^{-1} + B_{12}B_{22}^{-1}B_{21}.$$

This implies that

$$(R'R)_{11}^{-1} = (I_{ind}I_{ind})^{-1} + (R'R)_{12}^{-1}(R'R)_{22}(R'R)_{21}^{-1}.$$

Plugging this into Equation 3, we obtain

$$\alpha^f|D_{obs}, \sigma^2, \beta^f \sim N\left(\hat{\alpha}^f, \sigma^2(I_{ind}I_{ind})^{-1}\right),$$

or for each individual α_j

$$\alpha_j^f|D_{obs}, \sigma^2, \beta^f \sim N(\hat{\alpha}_f^f, \sigma^2/n_j).$$

2. Derivation of the Relationship Between $Var\left(\alpha_j^f|\boldsymbol{D_{obs}}, \sigma^2, \beta^f\right)$ and $Var\left(\alpha_j^r|\boldsymbol{D_{obs}}, \tau^2, \psi^2, \beta^r\right)$

We want to find the multiplicative factor x in

$$Var\left(\alpha_j^f|D_{obs}, \sigma^2, \beta^f\right) = x \cdot Var\left(\alpha_j^r|D_{obs}, \tau^2, \psi^2, \beta^r\right)$$

Given that the residual variances are equal in both models, that is, $\sigma^2 = \psi^2$, we have

$$\frac{\psi^2}{n_j} = x\frac{\psi^2\tau^2}{n_j\tau^2 + \psi^2}.$$

Solving for x, we obtain

$$
\begin{aligned}
x &= \frac{\psi^2 n_j \tau^2 + \psi^4}{n_j \psi^2 \tau^2} \\
&= 1 + \frac{\psi^2}{n_j \tau^2} \\
&= 1 + \frac{1}{n_j}\left(\frac{1-\rho}{\rho}\right).
\end{aligned}
$$

Acknowledgments

I am thankful to Stephanie Eckman, Jerry Reiter, and Hans Schneeweiß for very useful comments on earlier versions of this article. I also thank three anonymous referees and the editor for thoughtful suggestions which helped to improve the article.

Author's Note

This article uses data from the National Educational Panel Study (NEPS): Starting Cohort 4–9th Grade, doi:10.5157/NEPS:SC4:1.1.0. 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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the DFG grant DR 831/2-1.

References

Abowd, J. M., Stinson, M., & Benedetto, G. (2006). *Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project*. Technical report, Longitudinal Employer–Household Dynamics Program, U.S. Bureau of the Census, Washington, DC.

Andridge, R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, *53*, 57–74.

Aßmann, C., Steinhauer, H., & Zinn, S. (2012). Weighting the fifth and ninth grader cohort samples of the National Educational Panel Study, panel cohorts. Bamberg, Germany: National Educational Panel Study, University of Bamberg.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Zeitschrift für Erziehungswissenschaft (Special Issue 14): Education as a Lifelong Process. The German National Educational Panel Study (NEPS)*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Brown, E., Graham, J., Hawkins, J., Arthur, M., Baldwin, M., Oesterle, S., … Abbott, R. (2009). Design and analysis of the community youth development study longitudinal cohort sample. *Evaluation Review*, *33*, 311–324.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. London, England: Sage.

Carpenter, J., Goldstein, H., & Kenward, M. (2011). Realcom-impute software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*, 1–14.

Clark, N., Shah, S., Dodge, J., Thomas, L., Andridge, R., Awad, D., & Little, R. (2010). An evaluation of asthma interventions for preteen students. *Journal of School Health*, *80*, 80–87.

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. (2010). *The choice between fixed and random effects models: Some considerations for educational research*. (Forschungsinstitut zur Zukunft der Arbeit, No. 5287). Bristol, England: University of Bristol.

Drechsler, J. (2011a). Multiple imputation in practice—A case study using a complex German establishment survey. *Advances in Statistical Analysis*, *95*, 1–26.

Drechsler, J. (2011b). New data dissemination approaches in old Europe—Synthetic datasets for a German establishment survey. *Journal of Applied Statistics*, *39*, 243–265.

Drechsler, J. (2011c). *Synthetic datasets for statistical disclosure control—Theory and implementation*. New York, NY: Springer.

Drechsler, J., Bender, S., & Rässler, S. (2008). Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, *1*, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., & Zwick, T. (2008). A new approach for disclosure control in the IAB Establishment Panel—Multiple imputation for a better data access. *Advances in Statistical Analysis*, *92*, 439–458.

Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, *105*, 1347–1357.

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, *55*, 3232–3243.

Duchhardt, C., & Gerdes, A. (2013). *NEPS technical report for mathematics—Scaling results of starting cohort 4 in ninth grade* (NEPS working paper no. 22). Bamberg, Germany: National Educational Panel Study, University of Bamberg.

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

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). West Sussex, England: John Wiley & Sons.

Harville, D. A. (1999). *Matrix algebra from a statistician's perspective* (2nd ed.). New York, NY: Springer.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*, 224–232.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, *79*, 362–384.

Little, R. J. A., & Raghunathan, T. E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 617–622). Alexandria, VA: American Statistical Association.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika*. Advance online publication. doi: 10.1093/biomet/ast044

Liu, M., Taylor, J., & Belin, T. (1995). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. In *Proceedings of the Biometrics Section of the American Statistical Association* (pp. 142–147). Alexandria, VA: American Statistical Association.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, *9*, 538–558.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report—Scaling the data of the competence tests* (NEPS working paper no. 14). Bamberg, Germany: National Educational Panel Study, University of Bamberg.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, *27*, 85–96.

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology*, *32*, 143–150.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–590.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 20–34). Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for nonresponse in surveys*. New York, NY: John Wiley and Sons.

Schafer, J. L. (1997). *Imputation of missing covariates under a multivariate linear mixed model*. State College: The Pennsylvania State University.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 421–442.

Schenker, N., Raghunathan, T., & Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, *29*, 533–545.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*, 499–521.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introdcution to basic and advanced multilevel modeling*. London, England: Sage.

StataCorp. (2011). Accounting for clustering with mi impute. Retrieved February 21, 2014, from http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. Hox & J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Milton Park, England: Routledge.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45* 1–67.

von Maurice, J., Sixt, M., & Blossfeld, H.-P. (2011). The German National Educational Panel Study: Surveying a cohort of 9th graders in Germany (NEPS working paper no. 3). Bamberg, Germany: National Educational Panel Study, University of Bamberg.

Woolridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: The MIT Press.

Yucel, R. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modeling*, *11*, 351–370.

Yucel, R. M., He, Y., & Zaslavsky, A. M. (2008). Using calibration to improve rounding in imputation. *The American Statistician*, *62*, 125–129.

Zhao, J. H., & Schafer, J. L. (2013). *Pan: Multiple imputation for multivariate panel or clustered data*. R package version 0.9.

Zhou, X., & Reiter, J. P. (2010). A note on bayesian inference after multiple imputation. *The American Statistician*, *64*, 159–163.

Author

JÖRG DRECHSLER is a senior researcher at the Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany; e-mail: joerg.drechsler@iab.de. His research interests are in nonresponse and data confidentiality.

Manuscript received April 4, 2014
Revision received August 9, 2014
Accepted October 28, 2014