Received: 25 June 2021 | Revised: 17 June 2022 | Accepted: 20 June 2022

DOI: 10.1002/sim.9515

RESEARCH ARTICLE

Statistics in Medicine WILEY

Multiple imputation approaches for handling incomplete three-level data with time-varying cluster-memberships

Rushani Wijesuriya[1,2] | **Margarita Moreno-Betancur**[1,2] | **John Carlin**[1,2,3] | **Anurika Priyanjali De Silva**[3] | **Katherine Jane Lee**[1,2]

[1]Department of Pediatrics, Faculty of Medicine Dentistry and Health Sciences, The University of Melbourne, Melbourne, Victoria, Australia

[2]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Victoria, Australia

[3]Center for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia

Correspondence
Rushani Wijesuriya, Department of Pediatrics, Faculty of Medicine Dentistry and Health Sciences, The University of Melbourne, 50 Flemington road, Parkville, Victoria 3052, Australia.
Email: rushani.wijesuriya@mcri.edu.au

Funding information
Australian Research Council, Grant/Award Number: DE190101326; Victorian Government; National Health and Medical Research Council, Grant/Award Numbers: 1127984, APP1166023

Three-level data arising from repeated measures on individuals clustered within higher-level units are common in medical research. A complexity arises when individuals change clusters over time, resulting in a cross-classified data structure. Missing values in these studies are commonly handled via multiple imputation (MI). If the three-level, cross-classified structure is modeled in the analysis, it also needs to be accommodated in the imputation model to ensure valid results. While incomplete three-level data can be handled using various approaches within MI, the performance of these in the cross-classified data setting remains unclear. We conducted simulations under a range of scenarios to compare these approaches in the context of an acute-effects cross-classified random effects substantive model, which models the time-varying cluster membership via simple additive random effects. The simulation study was based on a case study in a longitudinal cohort of students clustered within schools. We evaluated methods that ignore the time-varying cluster memberships by taking the first or most common cluster for each individual; pragmatic extensions of single- and two-level MI approaches within the joint modeling (JM) and the fully conditional specification (FCS) frameworks, using dummy indicators (DI) and/or imputing repeated measures in wide format to account for the cross-classified structure; and a three-level FCS MI approach developed specifically for cross-classified data. Results indicated that the FCS implementations performed well in terms of bias and precision while JM approaches performed poorly. Under both frameworks approaches using the DI extension should be used with caution in the presence of sparse data.

KEYWORDS
clustered data, cross-classified data, missing data, multiple imputation, three-level data, time-varying cluster memberships

1 | INTRODUCTION

Clustered data structures are common in many clinical studies and occur due to observations being nested within groups, for example children within schools, and/or repeated measures within individuals (ie, longitudinal data).[1]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

These structures may be conceptualized hierarchically, with lower level units nested within higher level units, and are often analyzed using multilevel models.[2] There can be multiple levels of hierarchy within a given study. In our motivating example, the childhood to adolescence transition study (CATS) there are repeated measures collected at fixed time points (level 1) from students (level 2) belonging to different schools (level 3), resulting in a three-level data structure.[3]

Conventional multilevel models assume that the data are fully hierarchical, with each lower level unit (eg, each student) belonging to one and only one unit at the next higher level (eg, schools).[4,5] However, in longitudinal studies like the CATS, it may be that individuals move between clusters over the course of the study, for example in CATS children changed schools over time. This leads to a partially nonhierarchical, or cross-classified, structure where the repeated measures are clustered within individuals but the individuals are no longer clustered within the same higher-level cluster throughout the study period.[4] To model such data, cross-classified random effects models (CCREM) were developed, which explicitly model the time-varying higher-level cluster membership using time-varying random effects.[6,7] One such CCREM, is an "acute-effects" CCREM, in which the response of an individual at a fixed point in time is assumed to be only affected by the current higher-level cluster membership of the individual. While alternative more complex models have also been introduced in the literature for modeling time-varying cluster memberships (briefly detailed in Section 2), in this article we focus on an acute-effects CCREM as the motivating example in this article only involved exposure and outcome measurements at three time points, rendering the other models overly complex.

Missing data present challenges in many studies. In longitudinal studies like CATS, the repeated observations of individuals enhances the potential for participant fatigue and drop-out, thus increasing the risk of missing data. Multiple imputation (MI), initially proposed by Rubin, is widely used for handling missing data in longitudinal studies.[8] MI is a two-stage process. In the first stage, the missing values are imputed multiple times by sampling from an approximation to the posterior predictive distribution of the missing data given the observed data. In the second stage, the multiply imputed datasets are analyzed using the substantive analysis model of interest and the resulting inferences are combined using Rubin's rules.[8] The two commonly used frameworks for conducting MI when there are multiple incomplete variables are joint modeling (JM) and fully conditional specification (FCS). The JM approach imputes all variables with missing data simultaneously, by assuming a joint model for the incomplete variables conditioning on the complete variables. The most widely assumed joint model is a multivariate normal (MVN) model.[9,10] The FCS approach imputes the variables with missing values one at a time by cycling through a series of univariate imputation models for each incomplete variable conditioning on all the other variables.[11,12] It is now well established that the validity of the results from MI depends on the appropriate tailoring of the imputation model to the substantive analysis model to ensure congeniality between the two models.[13,14] In practical terms, this means all key features of the analysis model, such as any multilevel structures, need to be appropriately accommodated in the imputation model to ensure valid results.[15]

The standard implementations of both the JM and FCS MI approaches (referred to as single-level JM and single-level FCS) assume that the individual observations are independent. Using such approaches without accounting for the clustered structure can lead to biased results from a multilevel substantive analysis because the analysis and imputation models are uncongenial.[15-17] Two simple ways to incorporate the clustering within the single-level MI approaches is by extending them using dummy indicators (DIs) for cluster groups or imputing the repeated measures in wide format if measured at the same fixed time points for all individuals.[18-21] The DI approach can produce approximately unbiased estimates of the regression coefficients in the context of a random intercept substantive analysis model, when there are missing values in the outcome, and unbiased estimates of the regression coefficients at level 1, with missing values in outcome and covariates.[16,18] However, previous literature has shown some noteworthy limitations of the DI approach. First, the DI approach has been shown to overestimate the SEs of regression coefficients and the variance components if there are large amounts of missing data, the intra cluster correlation (ICC) is low, or the cluster sizes are small.[16,22] Second, the DI approach can overestimate higher-level variance components when there is missing outcome data,[16,22] and result in biased estimates of the higher-level regression coefficients when there is missing exposure data.[16,22] Finally, where data are missing completely in one or more clusters, fixed effects corresponding to the DIs of these clusters can face identifiability issues in the imputation model leading to underestimated higher-level variance components.[23,24] Imputing repeated measures in the wide format is theoretically a very flexible approach, as it does not impose a structure for the correlation matrix among repeated measures. Previous simulation studies have shown that the approach performs well in many contexts.[18,20,21,25] However, if there are a large number of variables, especially categorical variables, and/or multiple time points, imputing in wide format can lead to nonconvergence of the imputation model due to model over-fitting and/or collinearity.[20]

Recently, both the JM and FCS frameworks have been extended to impute multilevel data using linear mixed model (LMM)-based imputation models.[20,21,26,27] However, to date only a few implementations of MI approaches based on

three-level models are available, two in R through the "miceadds" and "mdmb" packages and one in the stand-alone Blimp software.[28,29] Alternatively, the single-level and the two-level MI approaches can also be extended for handling incomplete three-level data using the extensions described previously: a DI approach for the higher-level clusters and/or imputing the repeated measures in wide format.[18]

While all of the proposed extensions of single- and two-level MI approaches and the three-level MI approaches above, have been shown to produce valid inferences in handling incomplete purely hierarchical three-level data in simulations under various settings,[18,30] the implementation and comparability of these approaches in the context of cross-classified data structures is unclear. Yucel showed that using MI approaches that assume the data are purely hierarchical to handle incomplete nonhierarchical data, can result in biased estimates of the variance components, which can in turn lead to biased SEs for the regression coefficients.[31] However, MI approaches that can explicitly model cross-classified data structures in the imputation process are very limited. We are aware of only one MI approach that explicitly handles incomplete three-level cross-classified data, implemented in the R package "miceadds," which has not been evaluated in the MI literature.[29,32] Although the R package mdmb can potentially handle cross-classified data with an arbitrary number of levels,[32] there is limited documentation on how this can be implemented.

The scarcity of software and limited guidance may lead applied researchers to handle missing data in cross-classified data structures using convenient but improper approaches such as available case analysis (ACA) or MI procedures ignoring the time-varying cluster memberships, which may lead to biased results.[15] Given this gap in the literature, our aim was to evaluate approaches to handle missing data in this setting that are readily accessible or can be easily implemented within most statistical packages. We evaluated the following approaches: ignoring the time-varying cluster memberships by taking the first or the most common cluster (for comparison), pragmatic extensions of single- and two-level MI approaches within the joint modeling (JM) and the fully conditional specification (FCS) frameworks using dummy indicators and/or imputing repeated measures in wide format in a way that accounts for the cross-classified structure, and a three-level FCS MI approach developed specifically for cross-classified data. We did this in the context of an acute-effects CCREM substantive analysis model, using both a simulation study and a case study.

The organization of the article is as follows. We begin with a brief introduction to a case study in the CATS that motivated this research. The next section provides an overview of the MI approaches that are available for imputing incomplete cross-classified three-level data. We then describe a simulation study based on the CATS case study to evaluate the performance of these approaches before presenting the results from the various approaches applied to the CATS case study. Finally we conclude with a general discussion.

2 | THE MOTIVATING CASE STUDY

2.1 | Study background

The CATS is a longitudinal cohort study with a broad focus on examining the onset and course of educational, emotional, social, and behavioral problems in children as they transition from childhood to adolescence.[3] The participants were recruited from a stratified sample of 43 schools in metropolitan Victoria, Australia. All of the 2239 students in the third grade (8–9 years of age) of the participating schools were invited to participate and 1239 (55%) children were recruited to the study at wave 1 in 2012. Data have since been collected through annual follow-ups, using multiple sources including parent, teacher, and student self-report questionnaires, along with direct anthropometric measurements. By mid-2020, 8 waves of follow-up had been completed. All the participants were followed up for data collection irrespective of whether they moved schools during the study period. More details on the study can be found in the published protocol.[3]

2.2 | Target analysis of interest

The motivating research question aimed to estimate the effect of depressive symptoms (at waves 1, 2, and 3: exposure) on the subsequent academic performance of the students (at waves 2, 3, and 4: outcome), as measured by a numeracy rating provided by the school teacher. We limited our analysis to the first four waves of the study, up to the student's transition to high school, where the majority of students transferred to different schools. For simplicity, we omitted individuals who

TABLE 1 Description of variables of interest for the individual j at wave k, in the motivating case study in the CATS

Variable	Type	Grouping/range	Label
Child's sex	Categorical	0 = Female 1 = Male	sex_j
Child's age (wave 1) (years)	Continuous	Range [7–11]	age_{j1}
Standardized SES measured by the SEIFA IRSAD (wave 1)	Continuous	z-score	SES_{j1}
Teacher's numeracy rating (wave 1)	Continuous	Range [1–5]	$teacher\_score_{j1}$
[a]Teacher's numeracy rating (k = wave 2, 3, and 4)	Continuous	Range [1–5]	$teacher\_score_{jk}$
[b]Depressive symptoms (k = wave 1, 2, and 3)	Continuous	Range [0–8]	$depression_{jk}$
[c]Child behavior problems reported by SDQ (k = wave 1, 2, and 3)	Continuous	Range [0–40]	SDQ_{jk}

Abbreviations: IRSAD, index of relative socio-economic advantage and disadvantage; SDQ, strengths and difficulties questionnaire; SEIFA, socioeconomic index for areas; SES, socio-economic status.

[a]The rating provided by the classroom teacher assessing students mathematical skills which is measured on a 5 point Likert scale.

[b]A subset of 4 items (each ranging from 0 to 2) from the Short Mood and Feelings Questionnaire (SMFQ) was used to measure the depressive symptoms at each wave in the CATS.[3,33] The exposure measure (at each wave) was the total summary score of these four items.

[c]Derived from the first 4 subscales of the SDQ: emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems (each ranging from 0 to 10).[34] This variable was not included in the analysis but was included in the imputation model as an auxiliary variable to improve its performance.[35]

had missing information about which school they attended at any of the analysis waves (71/1239, 6%). To address this research question we used a CCREM, with random intercepts at both the school and individual level, to estimate the effect of the (time-varying) continuous exposure, the depressive symptoms score. In this analysis model, the random intercept at the school level was allowed to vary over time according to the varying school membership of the participant (highest level group membership).[7] The model also included adjustment for potential confounders measured at baseline (wave 1): teacher numeracy rating, sex, socio-economic status (SES), and age. Letting $i(j,k) \in \{1, \ldots, 181\}$ denote the school that individual j (j = 1, … ,1168) attended at wave k (k = 2, 3, 4), the analysis model of interest was:

$$\begin{aligned} \text{teacher\_score}_{jk} = \beta_0 + \beta_1 * \text{depression}_{j(k-1)} + \beta_2 * k + \beta_3 * \text{teacher\_score}_{j1} \\ + \beta_4 * \text{sex}_j + \beta_5 * \text{SES}_{j1} + \beta_6 * \text{age}_{j1} + \alpha_{i(j,k)} + \gamma_j + \varepsilon_{jk}, \end{aligned}$$

where $\alpha_{i(j,k)} \sim N\left(0, \sigma_3^2\right)$ is the school-level random effect for the school that individual j attended at wave k, $\gamma_j \sim N(0, \sigma_2^2)$ is the individual-level random effect and $\varepsilon_{jk} \sim N\left(0, \sigma_1^2\right)$ is the random error. A detailed description of the variables can be found in Table 1. Note that, although the child behavior problems reported by strengths and difficulties questionnaire (SDQ) in Table 1 is not a variable in the substantive analysis, it will be included in the imputation model as an auxiliary variable to improve its performance.[36]

The CCREM specified in (1) closely resembles a hierarchical three-level model but allows different school-level random effects at each time-point. It assumes that the response of an individual at a fixed point in time is affected only by the current higher-level group membership of the individual ("acute-effects") and that the effect of each school cluster is constant over time ("static-effects"). Other possible analysis models are cumulative-effects CCREM, where the response at a fixed point in time is allowed to be influenced by both current and past higher-level group memberships of the individual,[6,7] and dynamic-effects models that allow the effect of the groups to vary over time.[37] An alternative model, which is closely related to the cumulative-effects CCREM, is the multiple membership model (MMM). MMMs model time-varying cluster memberships by incorporating weights to represent the effect of multiple higher-level units on the lower level outcome. The weights are generally derived using the proportion of time an individual has spent in each cluster up until the time they are assessed .[6,25] Because these different models were overly complex for our motivating example and the currently functionality to fit cumulative- or dynamic-effects CCREMs is only available in dedicated software packages for multilevel modeling such as MLwiN, we restrict our attention to an acute- and static-effects CCREM, but return to these other models in the discussion.

In the CATS, data were missing for the (time-varying) exposure (depressive symptom score at wave 1, 2, and 3), the outcome (teacher rating score at wave 2, 3, and 4), baseline (wave 1) teacher rating scores and the auxiliary variable (child behavior problems reported by SDQ at wave 1, 2, and 3). The proportions of missing data in these variables by wave are given in Table 2.

TABLE 2 Missing data frequencies and proportions in the outcome, teacher rating score, the exposure, depressive symptom score, and the auxiliary variable,child behavior problems reported by SDQ, by wave in the substantive analysis ($n = 1168$)

Data collection wave	Teacher rating score frequency (%)	Depressive symptom score frequency (%)	Child behavior problems reported by the SDQ
1	87 (7%)	74 (6%)	31 (3%)
2	118 (10%)	82 (7%)	317 (27%)
3	103 (9%)	80 (7%)	290 (25%)
4	130 (11%)	119 (10%)	377 (32%)

3 | MI APPROACHES FOR HANDLING INCOMPLETE CROSS-CLASSIFIED THREE-LEVEL DATA

This section provides an overview of the MI approaches that we evaluated for imputing incomplete, cross-classified three-level data in the context of repeated measures at fixed time points nested within individuals clustered within higher-level clusters.

Ideally the time-varying cluster memberships would be incorporated in the imputation model by allowing the cluster membership to vary across waves within the imputation model (if feasible). Alternatively the mobility of individuals over time can be ignored forcing a hierarchical structure in the imputation model. Two common strategies used in applied research for analyzing cross-classified longitudinal data that ignore the mobility of participants are (1) assume that the participants stay in their initial higher-level cluster throughout the study (referred to as ***First-cluster*** approach) or (2) assume that the individuals remained in the higher-level cluster in which they spent most of their time in the study (referred to as ***Common-cluster*** approach).[38] These ad-hoc strategies ignoring mobility may also be used within the MI framework in handling incomplete cross-classified data and can be implemented using any MI approach that can handle three-level data.[18] Because these approaches can be implemented easily in software based on single-level extensions we consider these approaches in this article for comparison as described below. However theoretically these approaches are uncongenial with the analysis model and may thus lead to invalid inferences.

Below we describe the specific approaches that we consider in this manuscript.

3.1 | Single-level JM with DI indicators for the higher-level clusters and repeated measures imputed in wide format (JM-1L-DI-wide)

Single-level JM imputes all the variables with missing values simultaneously under a single-level joint distribution, which is usually assumed to be multivariate normal, for the incomplete variables. Incomplete categorical variables can be imputed directly as continuous variables followed by rounding[39] or assuming underlying normally distributed latent variables.[10]

With purely hierarchical data, the single-level JM approach can be used to impute three-level data by including as predictors a set of (I−1) DIs representing the higher-level cluster membership and imputing the repeated measures in wide format (i.e, with one row per individual and separate variables for each repeated measure).

Accounting for the cross-classified structure within this approach would require including a separate set of DIs for the higher-level clusters at each time point (wave). In our opinion this is not a sensible model, because this would mean that the DIs representing the school cluster membership at every wave will be included as predictors when imputing a repeated measure at a particular wave. Therefore we only consider the ***First-cluster*** and the ***Common-cluster*** approaches within ***JM-1L-DI-wide*** denoted ***JM-1L-DI-wide_f*** and ***JM-1L-DI-wide_c***, respectively. The ***JM-1L-DI-wide_f*** approach was implemented using the higher-level cluster membership at wave 2 (as the outcome is measured from wave 2).

3.2 | Single-level FCS with DI indicators for the higher-level clusters and repeated measures imputed in wide format (FCS-1L-DI-wide)

Similarly to ***JM-1L-DI-wide***, the single-level FCS approach can be used to handle incomplete (purely hierarchical) three-level data using DI indicators to represent the cluster membership of each individual, and imputing the repeated measures in wide format.[18] In the context of time-varying cluster membership, multiple sets of DIs can be included as predictors in the FCS procedure to represent the cluster membership at each time point, with the univariate imputation model specified for each incomplete repeated measure restricted to just include the set of DIs at that particular wave.

In our simulations we also consider the ad-hoc ***First-cluster*** and the ***Common-cluster*** strategies within this approach, denoted ***FCS-1L-DI-wide_f*** and ***FCS-1L-DI-wide_c***, respectively, as a comparison with the JM approaches and because these approaches are commonly used in practice.[40,41]

3.3 | Two-level FCS for the higher-level clusters with repeated measures imputed in wide format (FCS-2L-wide)

The two-level FCS extension introduced by van Buuren (2011) uses a series of univariate two-level LMMs to impute the missing values in the incomplete clustered variables. Under the ***FCS-2L-wide*** approach, repeated measures are included in wide format, treating each repeated measure of the same variable as a distinct variable in the imputation model. Univariate two-level LMMs are then specified for each incomplete repeated measure in turn with cluster-specific random effects to account for the correlation among individuals of the same higher-level cluster.

Similarly to ***FCS-1L-DI-wide***, time-varying cluster membership can be incorporated into ***FCS-2L-wide*** by including random effects in the imputation model to represent the cluster membership at the given time point being imputed.

A similar approach using a two-level imputation model to account for the higher-level clusters with repeated measures imputed in wide format, could be used to accommodate the three-level structure within the JM framework (referred to as ***JM-2L-wide***).[18,26] However to accommodate time-varying cluster memberships, this approach would require including random effects for the higher-level clusters for all time points (waves) to impute incomplete variables at a given time point, which is not feasible. It is also not clear how this would be implementable in practice. For these reasons we do not consider such an approach in this manuscript.

3.4 | Two-level JM for repeated measures with DI for higher-level clusters (JM-2L-DI)

As an alternative to ***JM-2L-wide***, the two-level JM approach can be used to allow for the clustering among repeated measures within an individual (imputing the data in long format that is, where each repeated measure is a separate row in the dataset), and then the correlation among individuals of the same higher level cluster and the time-varying cluster membership can be incorporated by including the relevant DI representing the cluster membership at each time-point.

3.5 | Two-level FCS for repeated measures with DI for higher-level clusters (FCS-2L-DI)

Analogously to ***JM-2L-DI***, the two-level FCS approach can be used to model the clustering of repeated measures within individuals using individual-specific random effects, with the relevant DI representing the cluster membership at each time-point.

3.6 | Three-level FCS (FCS-3L)

The three-level FCS approach is an extension of the two-level FCS approach by van Buuren (2011), where missing values are imputed in long format using a series of three-level LMMs for each incomplete variable. Under this approach the correlation among individuals within the same higher-level cluster and the clustering of repeated measures within individuals are modeled using cluster-specific and individual-specific random effects, respectively. Implementations of this approach are available in the standalone software Blimp,[28] and the R package miceadds (through the function ml.lmer).[29]

The three-level FCS approach can be used to handle cross-classified three-level data by imputing from a series of univariate CCREMs, thus allowing time-varying cluster-specific random effects for the higher-level clusters. This approach is available in the R package miceadds (through the function ml.lmer).[29,42]

For comparison we also applied the ***First-cluster*** and the ***Common-cluster*** approaches using a simple (noncross-classified) three-level FCS approach, since this approach uses a three-level hierarchical model similar to the substantive analysis and performed well in terms of bias and precision in our previous simulations in the context of purely hierarchical three-level data.[18] We denote these by ***FCS-3L _f*** and ***FCS-3L_c***, respectively.

The JM approach has also been extended to impute three-level data using a three-level MLMM.[43] Theoretically, this three-level JM approach could be extended to handle three-level cross-classified data, by imputing the missing values using a multivariate CCREM.[31] However, we are not aware of any implementations of this approach and hence this was not considered in this manuscript.

See Table 3 for a summary of all the MI approaches that we evaluated for handling cross-classified incomplete three-level data.

4 | SIMULATION STUDY

4.1 | Generation of complete data

In order to evaluate the performance of the MI approaches for imputing cross-classified three-level data described above, we conducted a simulation study based on the CATS. In this simulation study, data were generated as described below for individual j at wave k, with a total sample size of 1200 students. One thousand datasets were generated for a number of different scenarios (see Additional file 1: Table S1 for the parameter values used in the simulation study). For each simulation scenario, this limited the Monte Carlo SE for the coverage of nominal 95% confidence intervals to approximately 0.7%.[44] The steps used in the generation of the data for the simulation study were:

i 40 schools were generated at wave 1 and these were populated with varying numbers of students ranging from 8 to 66, distributed similarly to the CATS. To achieve this, the school cluster sizes $(8 \leq n_i \leq 66)$ were assumed to follow a truncated log-normal distribution and the cluster size for each school i was sampled randomly from this distribution. To set the total number of students across the 40 schools to 1200, the sampled cluster sizes were multiplied by a factor of $\frac{1200}{\sum_{i=1}^{n_i} n_i}$ and rounded to derive a scaled class size. If the total of these scaled class sizes was less than 1200, the deficit was added to the last school cluster, while if the total of scaled class size was higher than 1200, the excess was deducted from the last school cluster.

Child's age at wave 1 $\left(\text{age}_{j1}\right)$ was generated from a uniform distribution $U(a, b)$.

ii Child's sex $\left(\text{sex}_j\right)$ was generated by randomly assigning λ% of students to be female.
iii Child's standardized SES value at wave 1 $\left(\text{SES}_{j1}\right)$ was generated from a standard normal distribution, N(0,1).
iv The teacher's numeracy rating at wave 1 $\left(\text{teacher\_score}_{j1}\right)$ was generated from a linear regression model conditional on child's sex, child's age at wave 1 and child's SES:

$$\text{teacher\_score}_{j1} = \eta_0 + \eta_1 * \left[\text{sex}_j = 1\right] + \eta_2 * \text{age}_{j1} + \eta_3 * \text{SES}_{j1} + \psi_j, \tag{2}$$

where ψ_j are independently and identically (iid) distributed as $\psi_j \sim N(0, \sigma_\psi^2)$.

v To mimic the varying cluster membership seen in the CATS study, new schools were added at each of waves 2, 3, and 4. We considered two scenarios with different numbers of school clusters being added at each wave: the first where 50 new schools were added at each wave, and the second where 10 new schools were added at each wave. A proportion (5%) of students at each wave were then randomly selected to be moved to these newly generated schools, with equal numbers of students being assigned to each new school.

TABLE 3 Summary of the evaluated MI approaches for handling incomplete three-level cross-classified data

			How the two sources of clustering are handled		
MI approach	**Type**	**Software**	**Clustering due to higher level clusters**	**Clustering due to repeated measures**	**How the time-varying cluster memberships are handled**
JM-1L-DI-wide_f	Standard (single-level)	SAS, SPSS, Stata, Mplus, R	DI	Repeated measures imputed in wide format	Time-varying nature of cluster memberships is ignored and only the DI for cluster membership at the first time point is used for all time points
JM-1L-DI-wide_c					Time-varying nature of cluster memberships is ignored and only the DI for the most common cluster membership over the course of the study is used for all time points
FCS-1L-DI-wide	Standard (single-level)	SAS, SPSS, Stata, Mplus, R, Blimp	DI	Repeated measures imputed in wide format	Restricting the univariate imputation models specified for each incomplete repeated measure to just include the DI at that particular wave
FCS-1L-DI-wide_f					Time-varying nature of cluster memberships is ignored and only the DI for cluster membership at the first time point is used for all time points
FCS-1L-DI-wide_c					Time-varying nature of cluster memberships is ignored and only the DI for the most common cluster membership over the course of the study is used for all time points
FCS-2L-wide	Specialized for two levels	Mplus, R, Blimp	RE	Repeated measures imputed in wide format	Restricting the univariate imputation models specified for each incomplete repeated measure to just include the RE for the cluster at that particular wave
JM-2L-DI	Specialized for two levels	R, Realcom-impute, Stat-JR	DI	RE	Including the relevant DI representing the cluster membership at each time-point in long format
FCS-2L-DI	Specialized for two levels	R	DI	RE	
FCS-3L-ml.lmer	Specialized for three levels	R	RE	RE	Through time-varying REs for the clusters (ie, using a series of univariate CCREMs for imputation in FCS)
FCS-3L_f	Specialized for three levels	R, Blimp	RE	RE	Time-varying nature of cluster memberships is ignored and RE for the cluster group at first time-point is used at all time points
FCS-3L_c	Specialized for three levels	R, Blimp	RE	RE	Time-varying nature of cluster memberships is ignored and RE for the most common cluster membership over the course of the study is used at all time points

TABLE 4 Simulation scenarios

	Base-case	Smaller sample size	A higher number of waves	Small number of constant set of clusters
Number of school clusters	40	40	40	10
School cluster sizes	varying (ranging from 8–66)	varying (ranging from 8–66)	varying (ranging from 8–66)	constant size of 120
Total sample size	1200	300	1200	1200
Number of school clusters added at each wave	10	10	10	10
Number of waves of data collection	4	4	8	4

vi Child's depression status at waves 1, 2, and 3 $\left(\text{depression}_{jk}, k = 1, 2, 3\right)$ was generated using a CCREM conditional on child's age at wave 1, child's sex, teacher's numeracy rating at wave 1, child's SES and wave:

$$\text{depression}_{jk} = \delta_0 + \delta_1 * \text{age}_{j1} + \delta_2 * \left[\text{sex}_j = 1\right] + \delta_3 * \text{teacher\_score}_{j1} + \delta_4 * \text{SES}_{j1} + \delta_5 * k + \theta_{i(j,k)} + \omega_j + \varphi_{jk}, \quad (3)$$

where φ_{jk}, ω_j, and $\theta_{i(j,k)}$ are iid as $\varphi_{jk} \sim N\left(0, \sigma_\varphi^2\right)$, $\omega_j \sim N\left(0, \sigma_\omega^2\right)$, and $\theta_{i(j,k)} \sim N\left(0, \sigma_\theta^2\right)$, respectively.

vii The outcome, teacher numeracy rating score at waves 2, 3, and 4 $\left(\text{teacher\_score}_{jk}, k = 1, 2, 3\right)$ was generated using the target analysis model as per Equation (1).

viii Finally, child's behavioral problems at waves 1, 2, and 3 $(\text{SDQ}_{jk}, k = 1, 2, 3)$ an auxiliary variable associated with the exposure, was generated using a CCREM conditional on depression symptoms at waves 1, 2, and 3 and wave:

$$\text{SDQ}_{jk} = \gamma_0 + \gamma_1 * \text{depression}_{jk} + +\gamma_2 * k + \nu_{i(j.k)} + \xi_j + \varepsilon_{jk}, \quad (4)$$

where ε_{jk},$\nu_{i(j.k)}$ and ξ_j are iid as; $\varepsilon_{jk} \sim N\left(0, \sigma_\varepsilon^2\right)$ $\nu_{i(j.k)} \sim N\left(0, \sigma_\nu^2\right)$, and $\xi_j \sim N(0, \sigma_\xi^2)$, respectively.

In the simulation study, we considered two different values for the ICC at the school level (high = 0.15 and low = 0.01) and at the individual level (high = 0.5 and low = 0.2) in the data generating model for the outcome (i.e., Equation 1). In the context of a random effects model, ICC is defined as the ratio of variation at the cluster level (ie, the variance of the random effects) divided by the total variation.[16] Therefore, ICC at the school level is equal to the ratio of between-school cluster variation to the total variation, while the ICC at the individual level is the ratio of (total) between-individual variation to total variation (see Additional file 1: Equation S1). The values for the ICCs were chosen based on the estimated ICC values from the CATS (school-level ICC = 0.5 and individual-level ICC = 0.01) and multilevel literature (an ICC of 0.05 is common in cluster randomized trials and larger ICC values such as 0.2 are seen in repeated measures designs), as used in our previous simulation study based on the CATS.[18] We refer to the four simulation scenarios corresponding to four different combinations of ICC values as: High-high, High-low, Low-high, and Low-low. Under each of the ICC combinations, the variance components at the three levels in the final population data generating model were obtained by equating the total variance across all the three levels, that is, the sum of σ_1^2, σ_2^2, and σ_3^2, to unity (see Additional file 1: Table S2).

To increase generalizability, we also considered a few other simulation scenarios that could be relevant in other practical applications. Considering the simulation scenario similar to the CATS case study (ie, 40 school clusters with varying numbers of students clustered within each cluster at wave 1 totaling to 1200 students and 10 new schools being added at each additional wave of data collection) to be the "base-case", we considered three other slight variations to include (i)smaller number of clusters, (ii) higher number of waves, (iii) small samples, and (iv) a constant set of clusters as described in Table 4:

4.2 | Generation of missing data

Data were set to missing in depressive symptom score (the exposure of interest) at waves 1, 2, and 3 with 15%, 20%, and 30% missing at each wave, respectively. Missingness was generated according to two missing at random (MAR) mechanisms, labeled MAR-CATS and MAR-inflated, by drawing from a logistic regression model dependent on the teacher numeracy rating at the subsequent wave and SDQ measure at the concurrent wave, as shown below:

$$\text{logit}\left(P\left(\text{R\_}depression_{jk}=1\right)\right)=\zeta_{0k}+\zeta_1*\text{teacher\_score}_{j(k+1)}+\zeta_2*\text{SDQ}_{jk},$$

where R_depression_{jk} is an indicator variable which takes the value 0 if depression_{jk} is missing and 1 if depression_{jk} is observed.

For the MAR-CATS scenario we set $\zeta_1=\ln(2)$ and $\zeta_2=\ln(1.5)$, which represent the associations observed in the CATS. For the MAR-inflated scenario we set $\zeta_1=\ln(4)$ and $\zeta_2=\ln(3)$. The values of the intercepts ζ_{0k} were chosen by iteration so that the required proportions of missingness were achieved.

In the additional simulation scenario with 8 waves of data collection, approximately 10% of the data were set to missing in the depressive symptom scores at each wave from waves 1–7. This proportion was determined based on the average proportion of missing data across the 7 waves in the CATS data (see Table S3).

4.3 | MI model specification and evaluation

We applied the MI approaches ***JM-1L-DI-wide-f***, ***JM-1L-DI-wide-c***, ***FCS-1L-DI-wide_f***, ***FCS-1L-DI-wide_c***, ***JM-1L-DI-wide***, ***FCS-1L-DI-wide***, ***FCS-2L-wide***, ***JM-2L-DI***, ***FCS-2L-DI***, ***FCS-3L, FCS-3L _f***, and ***FCS-3L_c*** to data generated from the 19 simulated scenarios (2 different settings for the number of higher-level clusters added at each wave × 4 ICC combinations × 2 missing data mechanisms plus 3 additional scenarios with smaller sample size, larger number of waves and a constant number of higher-level clusters, respectively). As a benchmark, we also conducted an ACA where waves of an individual with missing values were excluded from the analysis.

All of the MI approaches were executed using the R software version 3.6.1. ***JM-1L-DI-wide*** and ***JM-2L-DI*** were implemented in the R package "jomo".[45] ***FCS-1L-DI-wide***, ***FCS-2L-wide***, and ***FCS-2L-DI*** were implemented in the R package "mice" using the functions norm and 2L.pan for the single and two-level models, respectively, chosen because they are the most widely applied functions among the functions available in R for MI.[46] ***FCS-3L, FCS-3L _f*** and ***FCS-3L- _c*** were implemented using the R package "miceadds" (version 3.10-28) using the function ml.lmer.

The imputation models specified for each MI approach included all the variables in the analysis model and child behavior problems at waves 1, 2, and 3 as auxiliary variables. For each simulated dataset, 20 imputations were generated for each of the MI approaches.[15] After visually examining the trace plots, the JM approaches in R were run with a burn-in of 1000 iterations and 100 between-imputation iterations (except in the scenario with a higher number of waves, where the between-imputation iterations were set to 10 to reduce the computational time) while the FCS approaches were run with a burn-in of 10 iterations.

Following imputation, the CCREM was fitted to each of the imputed datasets and the resulting inferences were combined using the mitml package in R.[47] The resulting estimates of the parameters of interest (regression coefficient for depressive symptoms $-\beta_1$, and the estimates of the variance components at levels 1, 2, and 3 ($\sigma_1^2, \sigma_2^2, \sigma_3^2$, respectively) were compared to the true values used to generate the data. The performance of the approaches for estimating the regression coefficients of interest were evaluated using the mean value of the estimate, the bias (the average difference between the true values used in generating the data and the estimates), the empirical SE (the average SD of the estimates), the model-based SE (the average of the estimated SEs of the estimates), and the coverage probability of the nominal 95% confidence interval (estimated as the proportion of replications in which the estimated interval contained the true value) across the 1000 replications.[48] For the variance component estimates, we report the bias, percentage bias defined as the bias relative to the true value and the empirical SE.

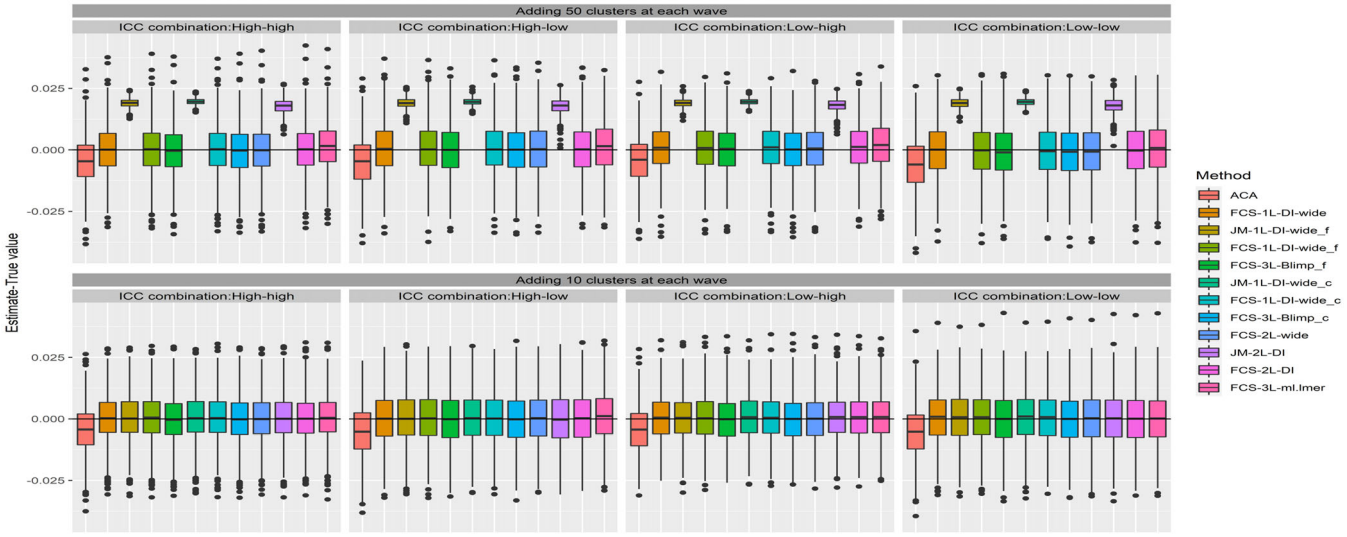

FIGURE 1 Distribution of the deviations of the estimated regression coefficient of interest from true value (β1,true value = −0.02) across the 1000 simulated datasets from available case analysis (ACA) and the 11 multiple imputation (MI) approaches under two scenarios of different number of higher level clusters and four ICC combinations when data are missing at random with dependencies based on CATS (MAR-CATS)

4.4 | Simulation study results

The sampling distribution of the deviations of the estimates from the true value of the regression coefficient of interest (β_1), across the 1000 replications for each approach for the different scenarios are shown in Figures 1 and 2. A detailed numerical summary of the estimated bias, standard errors and coverage is presented in Tables S4–S8 in Additional File 1. The comparative performance of the approaches was very similar across the different missing data mechanisms, so we focus mainly on the results from the MAR-CATS scenario, highlighting contrasts where they exist.

As expected, ACA resulted in large biases for β_1 (relative bias >10%) across all of the simulation scenarios with inadequate coverage (<93.6%).[48] In the scenario where there was a constant small number of higher-level clusters, all approaches performed well in estimating β_1. In the base-case scenario (large sample size, small number of follow up waves, and 10 clusters added at each wave), all of the JM approaches extended with the DIs for the school clusters (***JM-1L-DI-wide_c***, ***JM-1L-DI-wide_f***, ***JM-2L-DI)*** resulted in approximately unbiased estimates with appropriate nominal coverage. However, we observed large biases and severe undercoverage in the estimation of β_1 with all of the JM approaches when 50 clusters were added at each wave or when the sample size was small, and with ***JM-1L-DI-wide_c*** when there was a larger number of follow up waves. When the samples size was larger, while all FCS approaches resulted in approximately unbiased estimates of β_1 with coverage probabilities close to the nominal level across all simulation scenarios, slightly larger biases were observed for ***FCS-3L*** based approaches. This bias was more prominent in the MAR-inflated scenario when the ICC at level 2 was high. We also observed warnings indicating that some of the school cluster indicators were being dropped in the R software mice in the simulation scenarios with 50 clusters added at each wave or a larger number of waves in extensions of the FCS approaches using DIs (***FCS-1L-DI-wide_c***, ***FCS-1L-DI-wide_f*** and ***FCS-2L-DI***). When the sample size was small, somewhat large biases were also observed for the single level FCS extensions (***FCS-1L-DI-wide_c***, and ***FCS-1L-DI-wide_f)***.

Figures 3 and 4 display the empirical SE, with error bars showing the Monte Carlo SE, and the average model-based standard errors. With small sample sizes or large numbers of clusters added all of the JM approaches extended with the DIs for the school clusters (***JM-1L-DI-wide_c***, ***JM-1L-DI-wide_f***, ***JM-2L-DI)*** resulted in model-based standard errors that were larger than the empirical standard errors. The same was observed with a larger number of waves for ***JM-1L-DI-wide_c***. While all of the other MI approaches and ACA exhibited approximately comparable empirical and model-based standard errors under most scenarios, there were a few exceptions. Particularly of note, in the case with a small number of constant higher-level clusters (Figure 4C), all approaches (including the JM approaches above) resulted in somewhat different empirical errors to the average model-based standard errors under the Low -high ICC combination.

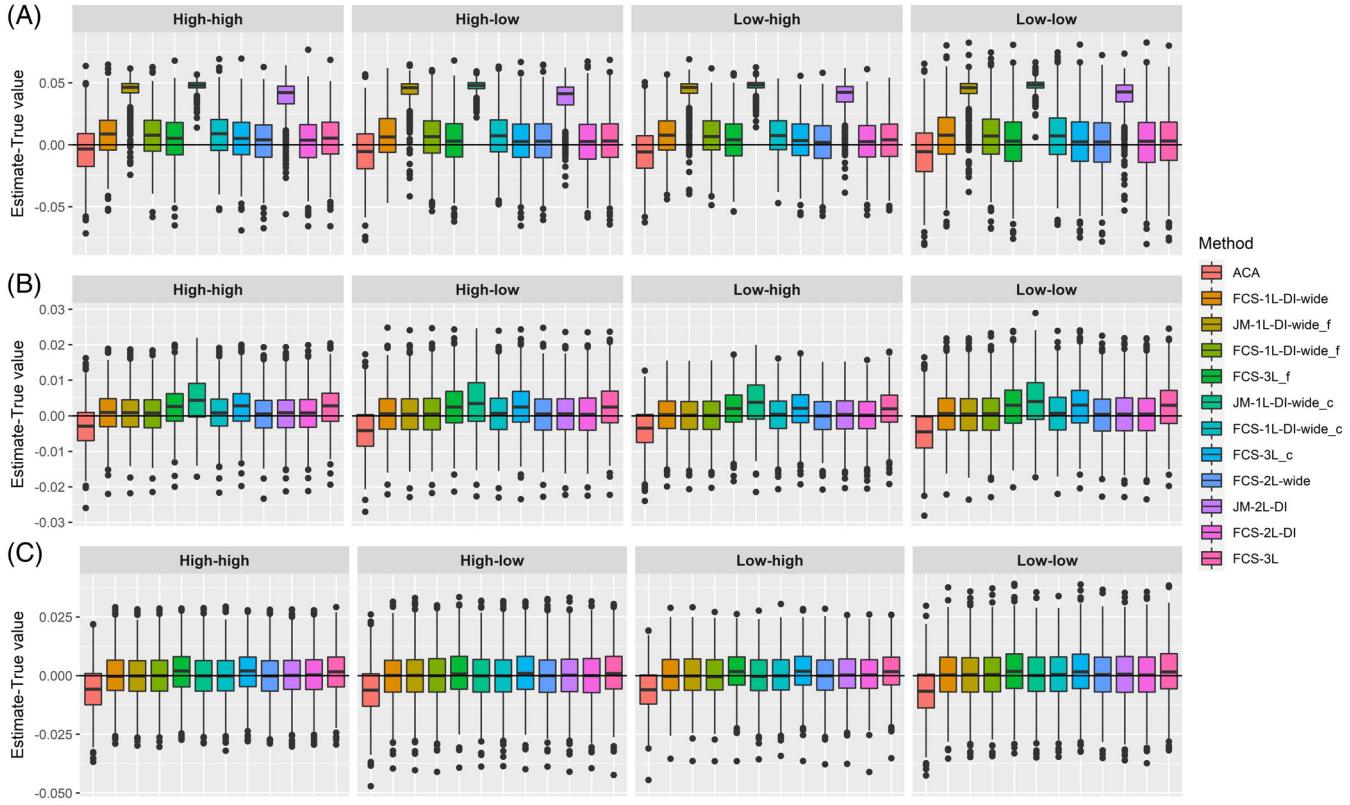

FIGURE 2 Distribution of deviations of the estimated regression coefficient of interest from true value across the 1000 simulated datasets for available case analysis (ACA) and 11 multiple imputation (MI) approaches with (A) a sample size of 300 (40 schools of varying school sizes at wave 1 and 10 new schools being added at each additional wave of data collection, and $\beta_1 = -0.05$) (B) 8 waves of data collection (40 school clusters of varying school sizes at wave 1 and 10 new schools being added at each additional wave of data collection, and $\beta_1 = -0.02$) (C) a small number of constant set of clusters (ie, no additional clusters being added at each wave, and $\beta_1 = -0.02$) when data are missing at random with inflated dependencies (MAR-inflated)

With small sample sizes (Figure 4A), FCS approaches also exhibited slightly different empirical standard errors and average model-based standard errors under the Low-high ICC combination.

The deviations from the true value in the variance component estimates at level 1, 2 and 3 for the scenario where 50 and 10 new clusters were added at each wave, are shown in Figures 5 and 6, respectively, while the results for the additional simulation scenarios are shown in the Figures S5–S7. All approaches showed similar performance with negligible bias (< 10% relative bias) for the variance components at level 1, 2, and 3 across the different simulation scenarios, with ACA showing comparatively larger biases.

5 | CASE STUDY ILLUSTRATION

All of the approaches were also applied to impute the incomplete data in the case study. In the CATS, missing values were observed in the outcome (teacher's numeracy rating at wave 2, 3, and 4), the exposure (depressive symptom scores at waves 1, 2, and 3), a time-fixed baseline variable (teacher numeracy rating scores at wave 1) and the auxiliary variable SDQ at waves 1, 2, and 3. Under the approaches that can accommodate the time-varying cluster memberships, all incomplete time-varying variables were imputed allowing the cluster membership to vary over time. The implementation details of each method in this multivariable missingness scenario and the corresponding code can be found here (https://3levelcrossclassified.netlify.app/). Similar to the simulation study we generated 20 imputations and included all the variables in the analysis model and an auxiliary variable-SDQ measure for child behavior problems in the imputation model for all approaches. Table 5 shows the results from the application to the CATS data (also see Additional file 1: Figure S8).

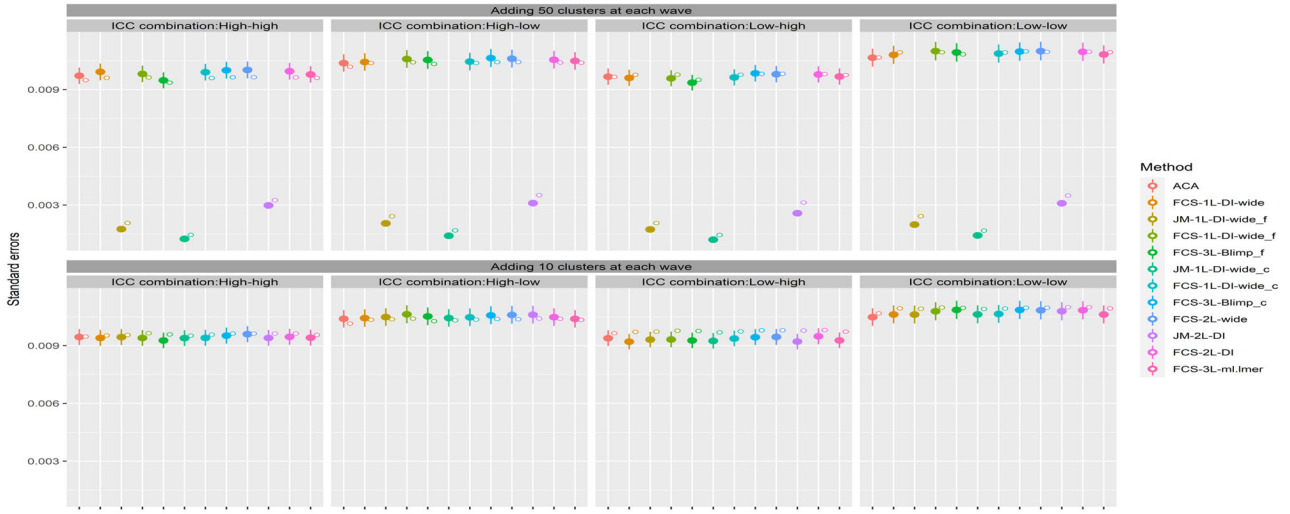

FIGURE 3 Empirical standard errors (filled circles with error bars showing ±1.96× Monte Carlo standard errors) and average model-based standard errors (hollow circles) for the regression coefficient of interest from 1000 replications, for available case analysis (ACA) and the 11 multiple imputation (MI) approaches under two scenarios of number of higher level clusters and four ICC combinations when data are missing at random with dependencies based on CATS (MAR-CATS)

All of the FCS approaches resulted in similar estimates for both the parameter of interest and the variance components. The exception was ***FCS-2L-DI*** , where the regression coefficient had a slightly larger SE and the level 3 and 2 variance components were somewhat larger than the other FCS approaches.

In contrast, all of the JM approaches resulted in estimates that were clearly implausible (very large estimates of the level 3 variance components, and a value of 0 for variance at level 2). A closer inspection of the observed and imputed values (see Additional File 1: Figures S9–S11) showed these three approaches generated many implausible imputed values for the incomplete variables which would lead to implausible results.

6 | DISCUSSION

In this article we compared a range of available approaches that can be used for handling incomplete, cross-classified, three-level data. We found that when the sample size was large, all of the extensions of single-level FCS approaches (including those that ignore the cross-classified structure), and extensions of two-level FCS approaches provided approximately unbiased estimates of the regression coefficients and the variance components across all simulation scenarios and in some cases performed better than the cross-classified three-level FCS approach. However, warnings were issued in the R package mice that some of the school clusters were being dropped for FCS approaches using the DI extension (***FCS-1L-DI-wide_c***, ***FCS-1L-DI-wide_f*** , and ***FCS-2L-DI)*** in the simulation scenarios where 50 new clusters were added at each wave and where there was a larger number of follow-up waves, as well as in the real data application. This indicates that the imputation models specified in mice were overparameterized, which is most likely due to sparsity of the data.[49] With small sample sizes, the single level FCS extensions (***FCS-1L-DI-wide***, ***FCS-1L-DI-wide_c***, and ***FCS-1L-DI-wide_f)*** performed poorly. The extensions of the single-level and two-level JM approaches (***JM-1L-DI-wide_f*** , ***JM-1L-DI-wide_c***, and ***JM-2L-DI***) performed well in the simulation scenario where only 10 new clusters were added at each wave, but performed very poorly when 50 new clusters were added at each wave or when the sample size was small. While all JM approaches performed well in terms of bias and precision in scenarios with a small number of constant clusters, ***JM-1L-DI-wide_c*** performed poorly when there was a larger number of follow-up waves. In line with the simulation results, all JM approaches extended with DI also produced implausible estimates in the real data illustration (where more than 30 schools were added at each wave with missing data in multiple variables).

The poor performance of the JM approaches using the DI approach in both the simulation and case study is likely due to sparse data, particularly where data are likely to be "systematically" missing, that is, missing for entire clusters. When data are systematically missing in clusters, fixed effects corresponding to the DIs for these clusters will be unidentifiable in the imputation model,[23,24] leading to nonconvergence of the JM imputation algorithm. In contrast, the FCS approach

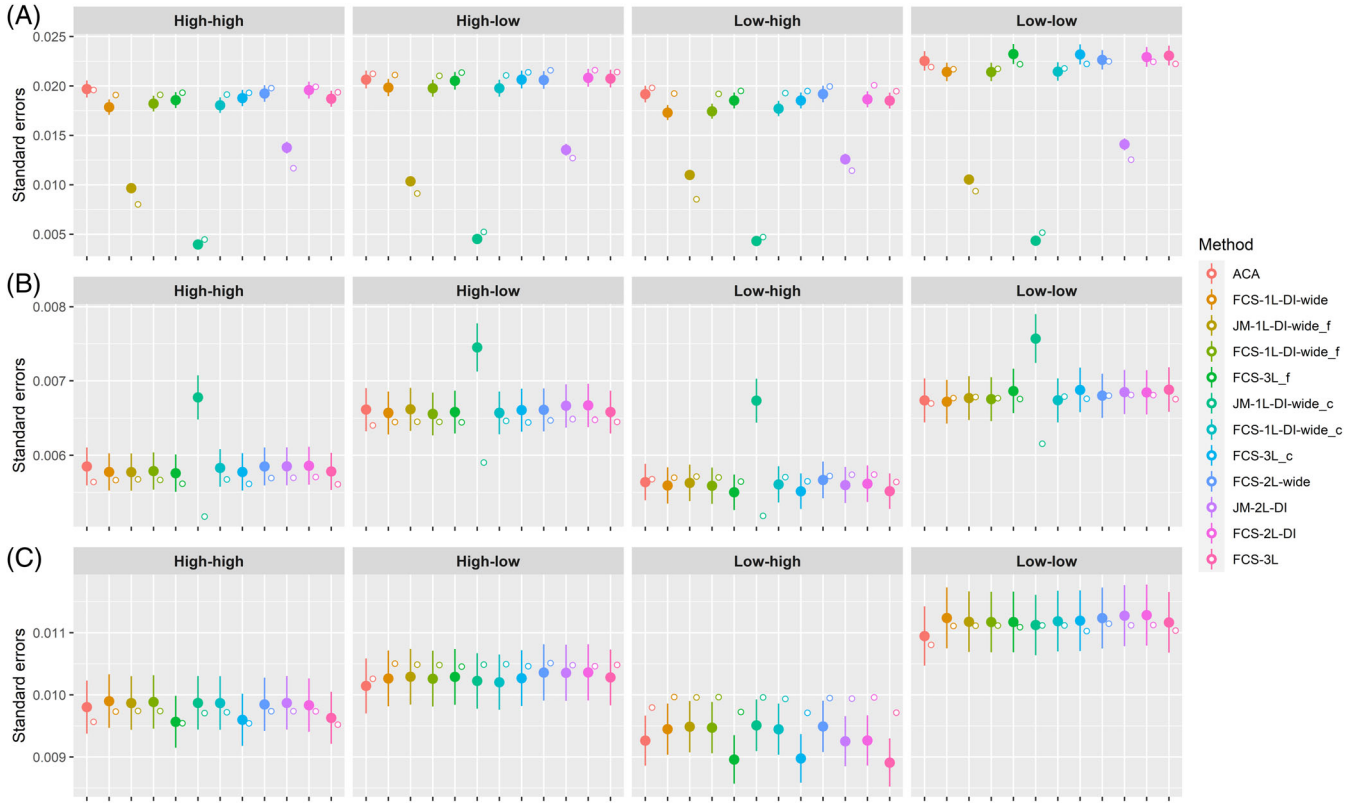

FIGURE 4 Empirical standard errors (filled circles with error bars showing ±1.96× Monte Carlo standard errors) and average model-based standard errors (hollow circles) for the regression coefficient of interest from 1000 replications, for available case analysis (ACA) and the 11 multiple imputation (MI) approaches with (A) a sample size of 300 (40 schools of varying school sizes at wave 1 and 10 new schools being added at each additional wave of data collection, and $\beta_1 = -0.05$) (B) 8 waves of data collection (40 school clusters of varying school sizes at wave 1 and 10 new schools being added at each additional wave of data collection, and $\beta_1 = -0.02$) (C) a small number of constant set of clusters (ie, no additional clusters being added at each wave, and $\beta_1 = -0.02$) when data are missing at random with inflated dependencies (MAR-inflated)

deals with this issue by dropping the DIs corresponding to the systematically missing clusters and hence will not face convergence issues. Although the FCS algorithm might converge, dropping DIs in this way could mean that the imputation model in FCS is misspecified if there were some cluster effects from the systematically missing clusters. Due to the likely issues of the DI approach with systematically missing clusters and the findings from previous literature that also indicate that the DI approach can produce inflated estimates of the variance components and the standard errors when there is a very low ICC, high missing data percentages or small cluster sizes,[16,50] we advise caution when using MI approaches extended with the DI approach. It should also be noted that while approaches that ignored the cross-classified structure performed well in our simulation study with large sample sizes, this was not the case with smaller sample sizes. Further research could investigate if there are other scenarios in practice where ensuring congeniality with the analysis model by accounting for the time-varying cluster memberships is necessary.

Our finding that pragmatic extensions of single and two-level FCS (***FCS-1L-DI-wide*** and ***FCS-2L-wide)*** can be used for handling incomplete three-level data with time-varying cluster memberships is important in practice because specialized approaches for imputing cross-classified three-level data can be challenging to implement due to limited availability in standard software packages and limited documentation, and there is also limited evaluation of these. In particular, many commonly used statistical software packages such as Stata and SPSS do not have MI approaches that impute using LMMs, leaving users of these software with the only option of using simple extensions of the single-level MI approaches for handling incomplete three-level data. It should be noted, however, that a disadvantage of the single-level extensions of the FCS approaches is that they are only feasible when data are collected at fixed time points for all individuals.

The simulation study in this article was based on a real study so we believe our results reflect what could be expected in practice in a similar setting. The simulation results also provided much-needed guidance on the use of MI approaches in the complex setting of incomplete cross-classified three-level data which commonly arises with longitudinal data.

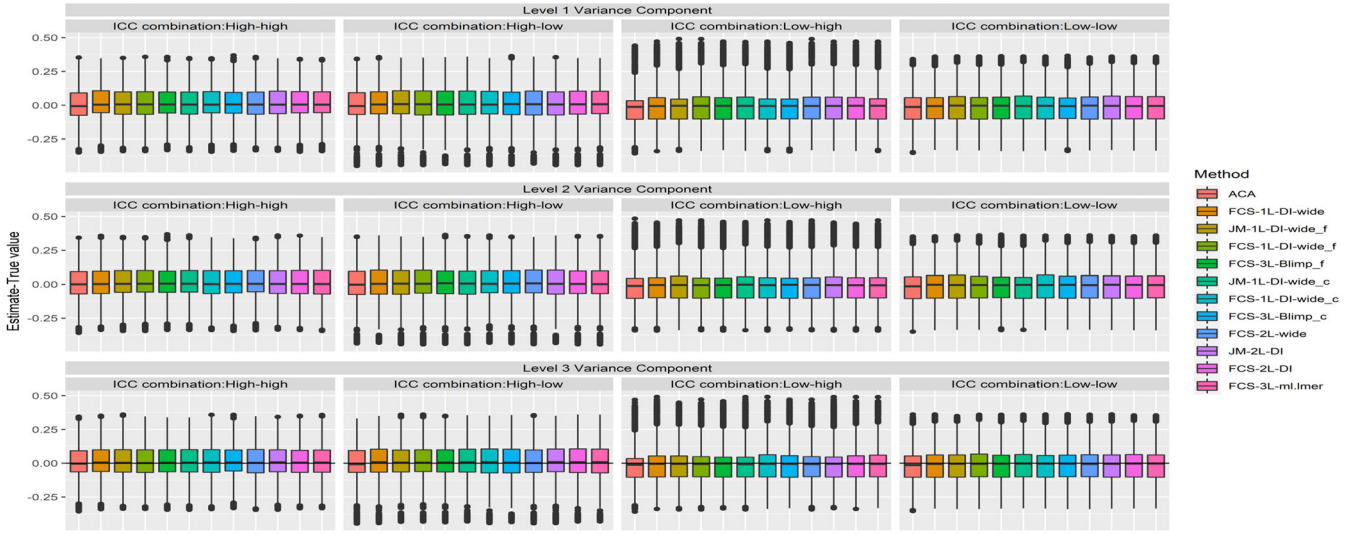

FIGURE 5 Distribution of the deviations of the variance component estimates from true values across the 1000 simulated datasets from available case analysis (ACA) and the 11 multiple imputation (MI) approaches under the simulation scenario with a higher number of higher-level clusters (addition of 50 new clusters at each wave) and data are missing at random with dependencies based on CATS (MAR-CATS)

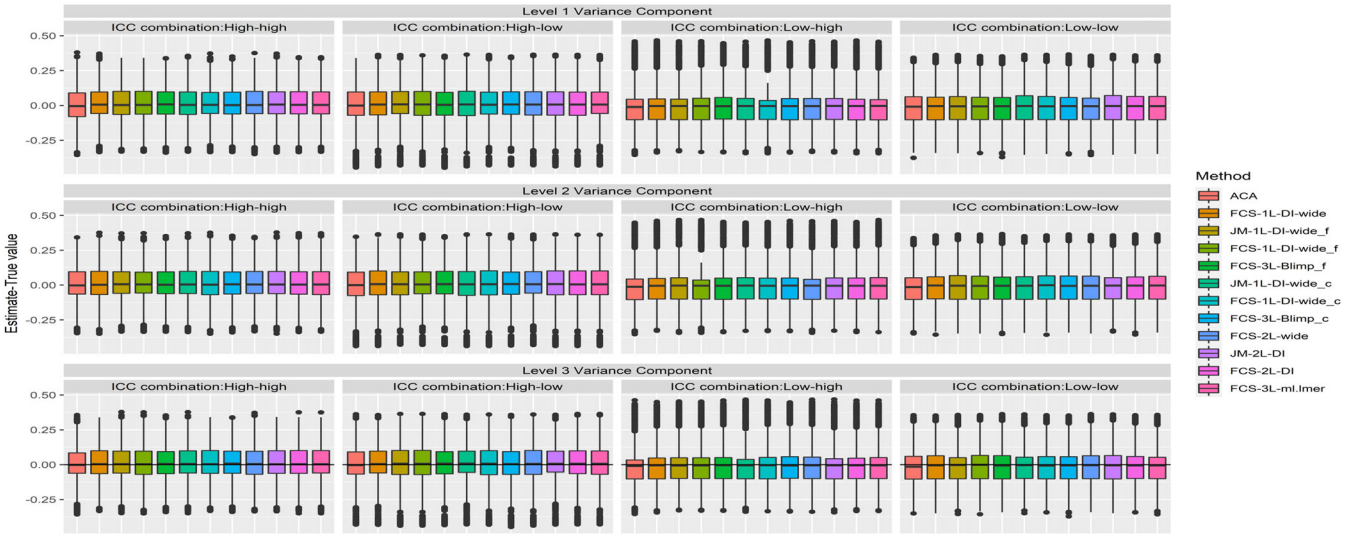

FIGURE 6 Distribution of the deviations of the variance component estimates from true values across the 1000 simulated datasets from available case analysis (ACA) and the 11 multiple imputation (MI) approaches under the simulation scenario with a lower number of higher-level clusters (addition of 10 new clusters at each wave) and data are missing at random with dependencies based on CATS (MAR-CATS)

However, it is always difficult to draw general recommendations from a single simulation study, as the performance of the methods may vary due to a range of factors that are at play when analyzing multilevel data, such as cluster sizes, number of clusters, ICCs, effect sizes and the variability and the number of the random effects, only a subset of which were evaluated in the current study. Therefore caution is required in generalizing these results to conditions outside those evaluated in our study.

In our simulations, the missingness was only imposed on a single level 1 continuous exposure, while in practice missing values often occur in multiple variables and at level 2 and/or level 3 and need to be handled simultaneously.[51] The results and recommendations may differ if the missingness is at a higher level and/or missing values also occur in outcome or confounding variables. Specifically, the approaches using the DI method have been shown to produce biased estimates of higher-level-variable regression coefficients when there are missing values in the outcome and the covariates. Therefore, future investigations of these approaches in a broader range of settings will be useful.

TABLE 5 Point estimate (and standard error) for the effect of depressive symptoms at the previous wave on the teacher numeracy scores, and point estimates for the variance components at levels 3, 2, and 1, from available case analysis (ACA) and MI approaches applied to the CATS data

Method		Regression coefficient estimate (SE)	Level 3 variance component	Level 2 variance component	Level 1 variance component
	ACA	−0.002 (0.007)	0.005	0.229	0.276
First-cluster approach (ignoring mobility)	***JM-1L-DI-wide_f***	***	***	***	***
	FCS-1L-DI-wide_f	−0.004 (0.007)	0.007	0.226	0.285
	FCS-3L_f	−0.005 (0.008)	0.006	0.194	0.307
Common-cluster approach	***JM-1L-DI-wide_c***	***	***	***	***
	FCS-1L-DI-wide_c	−0.003 (0.007)	0.008	0.229	0.287
	FCS-3L_c	−0.004 (0.007)	0.003	0.228	0.281
	FCS-1L-DI-wide	−0.004 (0.007)	0.009	0.227	0.286
	FCS-2L-wide	−0.004 (0.007)	0.009	0.227	0.286
	JM-2L-DI	***	***	***	***
	FCS-2L-DI	−0.002 (0.010)	0.029	0.301	0.285
	FCS-3L	−0.005 (0.007)	0.004	0.225	0.280

Note: ***All JM approaches produced implausible estimates and are therefore omitted from the table.

Our simulations were limited to the context of a static acute-effect CCREM and the MI approaches outlined may not be applicable in the context of alternative nonhierarchical models discussed previously such as MMMs or analysis models involving cumulative effects and/or dynamic effects. While limited theoretical work in the context of MMMs has been conducted,[31] developing novel MI implementations, as well as exploring whether the pragmatic extensions of the single- and two-level MI approaches can be used for handling incomplete data in such substantive contexts and comparing their performance, remains an avenue of future research. Future research could also investigate MI approaches in the context of other extensions to the substantive CCREM analysis model such as with random slopes, interaction effects, quadratic effects and more complex variance-covariance structures, where we may expect the performance of the approaches to differ, but this was beyond the scope of this study.[21,52,53] In our study we assumed that the cluster membership information is completely observed as the proportion of individuals who had missing cluster membership information at any of the analysis waves in our motivating example was very small (6%). In practice, however, this may not be the case, and cluster membership too may need to be imputed. van Buuren suggests that incomplete cluster membership may be handled using simple univariate imputation methods.[49] Alternatively, modeling approaches for analyzing cross-classified data with missing class identification as suggested by Hill and Goldstein could be used.[54]

In summary, our findings suggest that researchers may use the extensions of the single-level and two-level FCS approaches, ***FCS-1L-DI-wide*** and ***FCS-2L-wide***, evaluated in this study or the specialized three-level FCS approach for handling incomplete three-level cross-classified data. However, the specialized three-level approach may be the only option in settings with irregularly measured time points (ie, unbalanced data). Although FCS approaches, ***FCS-1L-DI-wide***, ***FCS-1L-DI-wide_c***, ***FCS-1L-DI-wide_f***, and ***FCS-2L-DI***, performed well with large sample sizes, these approaches can be problematic with small sample sizes and sparse data, and therefore should be used with caution. We recommend avoiding the use of extensions of the JM approaches with a large number of DIs as these approaches can lead to biased estimates of regression coefficients and variance components.

FUNDING INFORMATION

This work was supported by funding from the National Health and Medical Research Council: Career Development Fellowship ID#1127984 (Katherine Jane Lee) and project grant, ID#APP1166023. Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program. Margarita Moreno-Betancuris the recipient of an Australian Research Council Discovery Early Career Award (project number DE190101326) funded by the Australian Government. The funding bodies do not have any roles in the collection, analysis, interpretation and writing the manuscript.

DATA AVAILABILITY STATEMENT

The Childhood to Adolescence Transition Study (CATS) data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. The software code written for the simulations that support the findings of this study are openly available in a public GitHub repository at this link https://github.com/rushwije/MI3level_cross-classfied.

ACKNOWLEDGMENT

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

ORCID

Rushani Wijesuriya https://orcid.org/0000-0003-1023-4065
Margarita Moreno-Betancur https://orcid.org/0000-0002-8818-3125
Anurika Priyanjali De Silva https://orcid.org/0000-0003-0541-3202

REFERENCES

1. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Vol 998. Boca Raton: John Wiley & Sons; 2012.
2. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol 1. Thousand Oaks: Sage; 2002.
3. Mundy LK, Simmons JG, Allen NB, et al. Study protocol: the childhood to adolescence transition study (CATS). *BMC Pediatr*. 2013;13(1):160.
4. Luo W, Kwok O-m. The consequences of ignoring individuals' mobility in multilevel growth models: a Monte Carlo study. *J Educ Behav Stat*. 2012;37(1):31-56.
5. Luo W, Kwok O-m. The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivar Behav res*. 2009;44(2):182-212.
6. Cafri G, Hedeker D, Aarons GA. An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychol Methods*. 2015;20(4):407-421.
7. Raudenbush SW. A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *J Educ Stat*. 1993;18(4):321-349.
8. Rubin DB. *Multiple Imputation for Survey Nonresponse*. New York: Wiley; 1987.
9. Schafer JL. *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC Press; 1997.
10. Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Modell*. 2009;9(3):173-197.
11. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001;27(1):85-96.
12. van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. *J Stat Comput Simul*. 2006;76(12):1049-1064.
13. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9:538-558.
14. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med res*. 2015;24(4):462-487.
15. Carpenter J, Kenward M. *Multiple Imputation and Its Application*. Boca Raton: John Wiley & Sons; 2012.
16. Drechsler J. Multiple imputation of multilevel missing data—rigor versus simplicity. *J Educ Behav Stat*. 2015;40(1):69-95.
17. Black AC, Harel O, Betsy MCD. Missing data techniques for multilevel data: implications of model misspecification. *J Appl Stat*. 2011;38(9):1845-1865.
18. Wijesuriya R, Moreno-Betancur M, Carlin JB, Lee KJ. Evaluation of approaches for multiple imputation of three-level data. *BMC Med res Methodol*. 2020;20(1):1-15.
19. Kalaycioglu O, Copas A, King M, Omar RZ. A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *J R Stat Soc A Stat Soc*. 2016;179(3):683-706.
20. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol*. 2018;18(1):168.
21. Huque MH, Moreno-Betancur M, Quartagno M, Simpson JA, Carlin JB, Lee KJ. Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biom J*. 2019;62:444-466.
22. Lüdtke O, Robitzsch A, Grund S. Multiple imputation of missing data in multilevel designs: a comparison of different strategies. *Psychol Methods*. 2017;22(1):141-165.
23. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*. 2015;34(11):1841-1863.
24. Audigier V, White IR, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2018;33(2):160-183.
25. Wijesuriya RD. Evaluation of multiple imputation approaches for handling incomplete three-level data; 2021.

26. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *J Comput Graph Stat*. 2002;11(2):437-457.
27. van Buuren S. Multiple imputation of multilevel data. *Handbook of Advanced Multilevel Analysis.* New York: Routledge Taylor & Francis Group; 2011:173-196.
28. Keller B, Enders C. *Blimp User's Guide (Version 2.0)*. Los Angeles, CA; 2019.
29. Robitzsch A, Grund S, Henke T, Robitzsch MA. *Package 'Miceadds'*. Madison, WI: R Package; 2017.
30. Wijesuriya R, Moreno-Betancur M, Carlin JB, De Silva AP, Lee KJ. Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data. *Biom J*. 2021:1-22.
31. Yucel RM, Ding H, Uludag AK, Tomaskovic-Devey D. Multiple imputation in multiple classification and multiple-membership structures. Paper presented at: Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association; 2008.
32. Robitzsch A, Lüdtke O. mdmb: model based treatment of missing data (R package Version 1.0–18); 2018.
33. Angold A, Costello EJ, Worthman CM. Puberty and depression: the roles of age, pubertal status and pubertal timing. *Psycholog Med*. 1998;28(1):51-61.
34. Goodman R. Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry*. 2001;40(11):1337-1345.
35. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.
36. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.
37. Bauer DJ, Gottfredson NC, Dean D, Zucker RA. Analyzing repeated measures data on individuals nested within groups: accounting for dynamic group effects. *Psychol Methods*. 2013;18(1):1-14.
38. Grady MW, Beretvas SN. Incorporating student mobility in achievement growth modeling: a cross-classified multiple membership growth curve model. *Multivar Behav res*. 2010;45(3):393-419.
39. Galati JC, Seaton KA, Lee KJ, Simpson JA, Carlin JB. Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice. *J Stat Comput Simul*. 2014;84(4):798-811.
40. Mundy LK, Canterford L, Hoq M, et al. Electronic media use and academic performance in late childhood: a longitudinal study. *PLoS One*. 2020;15(9):e0237908.
41. Borschmann R, Mundy LK, Canterford L, et al. Self-harm in primary school-aged children: prospective cohort study. *PLoS One*. 2020;15(11):e0242802.
42. Grund S. Multiple imputation for three-level and cross-classified data. R Bloggers.
43. Charlton C, Michaelides D, Cameron B, et al. Stat-JR software; 2012.
44. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074-2102.
45. Quartagno M, Carpenter J. jomo: a package for multilevel joint modelling multiple imputation. R Package Version; 2016:2.2–0.
46. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2010;45:1-68.
47. Grund S, Robitzsch A, Luedtke O. mitml: tools for multiple imputation in multilevel modeling; 2017.
48. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279-4292.
49. van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton: Chapman &Hall/CRC Press; 2018.
50. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J*. 2011;53(1):57-74.
51. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data at level 2: a comparison of fully conditional and joint modeling in multilevel designs. *J Educ Behav Stat*. 2018;43(3):316-353.
52. Speidel M, Drechsler J, Sakshaug JW. Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. *Behav Res Methods*. 2018;50(5):1824-1840.
53. Wijesuriya R, Moreno-Betancur M, Carlin JB, De Silva AP, Lee KJ. Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data. 2021, Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data.
54. Hill PW, Goldstein H. Multilevel modeling of educational data with cross-classification and missing identification for units. *J Educ Behav Stat*. 1998;23(2):117-128.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wijesuriya R, Moreno-Betancur M, Carlin J, De Silva AP, Lee KJ. Multiple imputation approaches for handling incomplete three-level data with time-varying cluster-memberships. *Statistics in Medicine*. 2022;41(22):4385-4402. doi: 10.1002/sim.9515