Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data

Rushani Wijesuriya¹,² | Margarita Moreno-Betancur¹,² | John B. Carlin¹,²,³ | Anurika P. De Silva³ | Katherine J. Lee¹,²

¹ Department of Paediatrics, Faculty of Medicine Dentistry and Health Sciences, The University of Melbourne, Parkville, Victoria, Australia
² Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia
³ Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia

Abstract
Three-level data structures arising from repeated measures on individuals clustered within larger units are common in health research studies. Missing data are prominent in such studies and are often handled via multiple imputation (MI). Although several MI approaches can be used to account for the three-level structure, including adaptations to single- and two-level approaches, when the substantive analysis model includes interactions or quadratic effects, these too need to be accommodated in the imputation model. In such analyses, substantive model compatible (SMC) MI has shown great promise in the context of single-level data. Although there have been recent developments in multilevel SMC MI, to date only one approach that explicitly handles incomplete three-level data is available. Alternatively, researchers can use pragmatic adaptations to single- and two-level MI approaches, or two-level SMC-MI approaches. We describe the available approaches and evaluate them via simulations in the context of three three-level random effects analysis models involving an interaction between the incomplete time-varying exposure and time, an interaction between the time-varying exposure and an incomplete time-fixed confounder, or a quadratic effect of the exposure. Results showed that all approaches considered performed well in terms of bias and precision when the target analysis involved an interaction with time, but the three-level SMC MI approach performed best when the target analysis involved an interaction between the time-varying exposure and an incomplete time-fixed confounder, or a quadratic effect of the exposure. We illustrate the methods using data from the Childhood to Adolescence Transition Study.

KEYWORDS
congeniality, interactions, multiple imputation, non-linearities, substantive model compatible, three-level data
INTRODUCTION

Three-level data structures are common in observational data. For example, in life-course epidemiological studies, clustering of repeated measures from individuals (i.e., longitudinal data) who are further grouped within larger higher-level clusters, or where there are multiple layers of clustering such as participants within classes within schools. The Childhood to Adolescence Transition Study (CATS) is one such study where a cohort of students recruited from schools in Victoria, Australia, is followed up at multiple waves on a range of measures (Mundy et al., 2013). Missing data are a common challenge in many studies, but can be particularly problematic in longitudinal studies such as the CATS where there are multiple waves of data collection.

Multiple imputation (MI) is a widely used approach for handling missing data (Rezvan et al., 2015). MI is a procedure whereby missing values are imputed multiple times using an imputation model based on the observed data, followed by analyzing each imputed dataset using the intended substantive analysis model, and then appropriately pooling the resulting inferences across the multiply imputed datasets (Rubin, 1978). MI approaches impute the missing values from the conditional distribution of missing data given the observed data in an appropriate model for the observed and missing values. There are two common parametric model-based approaches to MI for imputing missing data in multiple variables: (i) joint modeling (JM), which imputes all partially observed variables simultaneously by specifying a joint model for the partially observed variables, conditional on any fully observed variables (Schafer, 1997); and (ii) fully conditional specification (FCS, also known as multiple imputation via chained equations) that imputes by iteratively sampling from a series of univariate conditional models for each partially observed variable conditional on all the other variables (Raghunathan et al., 2001; van Buuren et al., 2006). It has been shown that for MI to generate valid estimates of the substantive model parameters, the imputation model needs to be congenial with the analysis model, that is, it should preserve all the key features of the analysis such as nonlinear relationships, interactions, and multilevel features (Meng, 1994; Schafer, 2003).

In the case of a multilevel substantive analysis model, failure to incorporate the multilevel structure in the imputation model can lead to biased estimates of the regression coefficient and their standard errors (SEs), and also may severely bias estimates of the variance components (Black et al., 2011; Carpenter & Kenward, 2012). In the context of three-level data such as the CATS, there are a number of approaches available for incorporating the multilevel structure in the imputation model: imputation based on three-level models or extensions of single-level or two-level imputation approaches using dummy indicators (DIs) to represent the higher level cluster membership and/or by including repeated measures as distinct variables (arranged in “wide format”). Although the DI extension for handling multilevel data has been shown to produce biased estimates of substantive model parameters in some cases (Drechsler, 2015; Lüdtke et al., 2017; Speidel et al., 2018), in our previous work we found that these pragmatic adaptations to the single-level and two-level MI methods, and the three-level MI methods, all produced valid inferences in the context of a multilevel, random intercepts (at both levels) analysis model (Wijesuriya et al., 2020). However, similar comparisons of these approaches under more complex substantive analysis models such as when the analysis model includes interactions or quadratic effects involving incomplete covariates are lacking (van Buuren, 2018).

Several ad hoc ways of handling the interaction and/or quadratic effects using the standard MI approaches have also been proposed and evaluated in the literature, although the majority of studies have focused on single-level models (Allison, 2002; Seaman et al., 2012; Von Hippel, 2009). One strategy is to “impute-then-transform” also known as “passive imputation,” where the nonlinear or interaction terms are ignored in the imputation process and passively derived post-imputation. Alternatively, within the FCS framework, “passive imputation” can be incorporated at each iteration, so that the interactions or nonlinear terms are updated to incorporate the most recent imputations and can be used to impute the other incomplete variables (Royston, 2004). Another approach is to “transform-then-impute” the nonlinear or interaction term, where the nonlinear or interaction term is created prior to imputation and treated as “just another variable” (JAV) in the imputation model (Carpenter & Kenward, 2012; Seaman et al., 2012; Von Hippel, 2009; White et al., 2011). All of these approaches have been shown to produce biased regression coefficient estimates when the substantive model includes interaction or nonlinear effects when data are missing at random (MAR) (Grund et al., 2018; Seaman et al., 2012; van Buuren, 2018; Vink & van Buuren, 2013). It has also been suggested to use a reverse imputation strategy, where if the analysis model includes an interaction, the imputation model for the incomplete variable(s) includes an interaction between the outcome and the variables involved in the interaction (Grund et al., 2018; Tilling et al., 2016). However, analytical derivations of the conditional distributions of the incomplete covariates for several analysis models involving interactions and nonlinear terms clearly show that the conditional distributions for the incomplete covariates specified by all of the above ad hoc approaches are not compatible with those implied by the substantive analysis model (Enders et al., 2019; Kim et al., 2015; Lüdtke et al., 2019).
Bartlett et al. (2015) introduced substantive model compatible (SMC) MI that involves generating imputed values from appropriately specified conditional distributions of incomplete covariates for a given substantive model (Bartlett et al., 2015; Carpenter & Kenward, 2012; Enders et al., 2019; Goldstein et al., 2014). This is achieved by specifying the imputation model using a decomposition of the joint distribution of all variables into a conditional distribution for the outcome given covariates, one which is set to align with the substantive analysis model, and a joint distribution for the covariates (Bartlett et al., 2015; Lüdtke et al., 2019). SMC-MI can be implemented using joint modelling (SMC-JM), fully conditional specification (SMC-FCS) or sequential modelling (SMC-SM), a variation of the JM approach that factorizes the joint distribution of the variables into a sequence of conditional distributions (Lüdtke et al., 2019). In the context of single-level data, all these approaches have been shown to perform better than ad hoc extensions such as JAV and the different passive imputation approaches when the analysis model includes an interaction or a non-linear term (Enders et al., 2019; Kim et al., 2015; Lüdtke et al., 2019; Zhang & Wang, 2017). Although SMC-MI has been extended to the context of multilevel data, the extensions and their evaluations are limited to two-level data with no guidance for the settings of three-level data (Hayes, 2019; Keller, 2019). We are aware of only one SMC-MI approach that has been specifically extended for handling incomplete three-level data, which is implemented in the Blimp software (Keller & Enders, 2019) although this approach has not been fully evaluated and has not been compared with pragmatic adaptations such as JAV and passive imputation in single- and two-level MI approaches, or two-level SMC-MI approaches in the context of interactions or non-linear terms with three-level data. Therefore, the aim of the current paper is to compare the available approaches for imputing three-level data in the context of substantive analysis models involving interactions or non-linear terms commonly used in longitudinal data settings, using both simulations and a case study. We focus on multilevel data resulting from repeated measures with follow-ups at fixed intervals of time within an individual where there is clustering among individuals as in the CATS.

The remainder of the paper is structured as follows. Section 2 describes the case study that motivated this research and the substantive analysis models of interest. Section 3 provides an overview of the approaches that can be used to impute incomplete three-level data and how they can be used to accommodate the interactions or quadratic effects of covariates. In Section 4, we describe a simulation study based on the CATS case study in which we evaluate the available approaches, comparing their performance in terms of bias and precision. In Section 5, we apply these approaches to the analysis of the case study. We conclude with a general discussion in Section 6.

2  THE MOTIVATING EXAMPLE: THE CHILDHOOD TO ADOLESCENCE TRANSITION STUDY

The motivation for this study came from the CATS, a longitudinal cohort study that focuses on educational, emotional, social, and behavioral changes in children as they transition from puberty to adolescence (Mundy et al., 2013). The participants were recruited from a stratified random sample of 43 schools in Melbourne, Australia. All grade 3 students (approximately 8–9 years of age) enrolled in these schools were invited to participate. Of the 2239 invited students, 1239 (54%) children with informed parental/guardian consent were recruited into the study at wave 1 (2012). Data were collected through annual follow-ups using parent, teacher, and student self-report questionnaires, with currently seven waves of follow-up available for analysis. There is also linkage with the Victorian Curriculum and Assessment Authority to obtain National Assessment Programme – Literacy and Numeracy (NAPLAN) results. NAPLAN is a nationwide test administered to students in grades 3, 5, 7, and 9 (approximate ages 8–9, 10–11, 12–13, and 14–15 years), which assesses the student’s academic performance on four domains: reading, writing, numeracy, and language conventions. The detailed study protocol can be found elsewhere (Mundy et al., 2013).

2.1  Substantive analysis models

The motivating example for our study aimed to estimate the effect of early depressive symptoms (at waves 2, 4, and 6) on the academic performance of the students at the subsequent wave (waves 3, 5, and 7) as measured by NAPLAN numeracy scores with adjustment for potential confounders measured at baseline (wave 1): child’s NAPLAN numeracy scores, sex, socio-economic status (SES), and age (Mundy et al., 2017). With complete data, the standard modeling framework for estimating these effects would be an LMM. We focus on three LMMs involving interactions and quadratic terms, as defined by Equations (1)–(3), which are typical of the models that are of interest to researchers in longitudinal data settings. In
TABLE 1  Summary of variables of interest, in the motivating case study in the CATS

| Variable                                      | Type          | Grouping / Range | Label              |
|-----------------------------------------------|---------------|------------------|--------------------|
| Child’s sex                                   | Categorical   | 0 = Female 1 = Male | sex_{ij}           |
| Child’s age (wave 1) (years)                  | Continuous    | Range [7–11]     | age_{ij}           |
| Standardized SES measured by the SEIFA IRSAD (wave 1) | Continuous   | z-score          | SES_z_{ij}         |
| Standardized NAPLAN numeracy score (wave 1)   | Continuous    | z-score          | NAPLAN_z_{ij}      |
| Standardized NAPLAN numeracy score (k = wave 3, 5, and 7) | Continuous   | z-score          | NAPLAN_z_{ijk}     |
| Depressivesymptoms (k = waves 2, 4, and 6)    | Continuous    | Range [0–8]      | depression_{ij(k−1)} |
| Overall child behavior reported by SDQ (k = waves 2, 4, and 6) | Continuous   | Range [0–40]     | SDQ_i_{ijk(k−1)}   |

Abbreviations: IRSAD, Index of Relative Socio-economic Advantage; NAPLAN, National Assessment Program - Literacy and Numeracy; SDQ, Strengths and Difficulties Questionnaire; SEIFA: Socioeconomic Index for Areas, SES: Socio-Economic Status.

A subset of four 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 study (Angold et al., 1998; Mundy et al., 2013). These items are summed and centered at the mean to give a depressive symptoms score.

For measuring the overall child behavior, a total behavioral difficulties score is derived by summing the first four subscales of the Strengths and Difficulties Questionnaire (SDQ): emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems (each ranging from 0 to 10) (Goodman, 2001).

these model specifications, i denotes the school (i = 1, ..., 43), j denotes the individual (j = 1, ..., 1239), and k denotes the wave of data collection (k = 3, 5, 7), with ε_{ijk} a random error distributed as N (0, σ^2_i) and school and individual-level random effects α_{0i} ∼ N(0, σ^2_3) and α_{0ij} ∼ N(0, σ^2_2), respectively.

A detailed description of the variables in these models is provided in Table 1.

The analysis models of interest are:

1. A random intercept model with an interaction between the time-varying exposure and time

   \[ NAPLAN_{z_{ijk}} = \beta_0 + \beta_1 \ast depression_{ij(k−1)} + \beta_2 \ast wave_{ijk} + \beta_3 \ast depression_{ij(k−1)} \ast wave_{ijk} + \beta_4 \ast NAPLAN_{z_{ij1}} + \beta_5 \ast sex_{ij} + \beta_6 \ast SES_{ij1} + \beta_7 \ast age_{ij1} + \alpha_{0i} + \alpha_{0ij} + \epsilon_{ijk}. \]  

   Such a model allows the association between the exposure and the outcome to vary with time, with \( \beta_1 \) representing the effect of depressive symptoms on the standardized NAPLAN scores at time = 0, and \( \beta_3 \) representing the change in the effect of depressive symptoms on the standardized NAPLAN scores between two consecutive waves.

2. A random intercept model with an interaction between the time-varying exposure and a time-fixed baseline variable

   \[ NAPLAN_{z_{ijk}} = \beta_0 + \beta_1 \ast depression_{ij(k−1)} + \beta_2 \ast wave_{ijk} + \beta_3 \ast depression_{ij(k−1)} \ast SES_{ij1} + \beta_4 \ast NAPLAN_{z_{ij1}} + \beta_5 \ast sex_{ij} + \beta_6 \ast SES_{ij1} + \beta_7 \ast age_{ij1} + \alpha_{0i} + \alpha_{0ij} + \epsilon_{ijk}. \]  

   This model allows the association between the exposure and the outcome to vary with the level 2 variable, in this case SES. Here, \( \beta_1 \) represents the effect of depressive symptoms on the standardized NAPLAN scores when SES value is equal to the mean SES, \( \beta_3 \) represents the change in the effect of depressive symptoms on the standardized NAPLAN score for a 1 standard deviation change in the SES.

3. A random intercept model with a quadratic effect of the time-varying exposure

   \[ NAPLAN_{z_{ijk}} = \beta_0 + \beta_1 \ast depression_{ij(k−1)} + \beta_2 \ast wave_{ijk} + \beta_3 \ast depression^2_{ij(k−1)} + \beta_4 \ast NAPLAN_{z_{ij1}} + \beta_5 \ast sex_{ij} + \beta_6 \ast SES_{ij1} + \beta_7 \ast age_{ij1} + \alpha_{0i} + \alpha_{0ij} + \epsilon_{ijk}. \]  

   This model implies that the association between the exposure and the outcome is nonlinear, the shape of which is characterized by \( \beta_1 \) and \( \beta_3 \).

In the CATS, data were missing for the (time-varying) exposure, depressive symptom score, and the outcome, NAPLAN numeracy scores. NAPLAN numeracy scores were missing for 15% (184/1239) of individuals at wave 1, 16% (198/1239) at wave 3, 21% (264/1239) at wave 5, and 30% (366/1239) at wave 7. Depressive symptom scores were missing for 11% (137/1239) of individuals at wave 2, 14% (173/1239) at wave 4, and 21% (249/1239) at wave 6.
3 OVERVIEW OF POSSIBLE MI APPROACHES FOR HANDLING INCOMPLETE THREE-LEVEL DATA IN THE CONTEXT OF INTERACTIONS AND/OR QUADRATIC TERMS

In this section, we outline the MI approaches that can be used to impute incomplete three-level data in the context of the analysis models outlined in Section 2, describing how each approach handles the nonlinear or interaction term as well as the two sources of clustering: in our context the correlation among individuals belonging to the same school and the correlation among the repeated measures of an individual.

(i) Single-level JM with DI indicators for the higher level clusters and repeated measures imputed in wide format (JM-IL-DI-wide)

Single-level JM, popularized by Schafer (1997), imputes the missing values by assuming a joint distribution for the incomplete variables. Imputations for the missing values are drawn from the joint posterior predictive distribution of the missing data given the observed data (Schafer, 1997). Commonly the joint distribution is assumed to be multivariate normal (MVN) (including for categorical variables), where the incomplete variables are included as outcomes and the complete variables are included as predictors in the imputation model. This approach can be implemented in most statistical software. A slight variation to the standard MVN approach is where incomplete categorical variables are modeled directly as a part of the joint model by modeling them as normally distributed latent variables (Goldstein et al., 2009). This approach has been implemented in the R package “jomo” (Quartagno & Carpenter, 2016).

A simple way of adapting the JM approach to handle three-level data with repeated measures and clustering is to include a set of DIs (a total of $I - 1$ DIs for $I$ higher-level clusters) representing the cluster membership of each individual. This estimates a separate intercept/fixed effect for each cluster (Lüdtke et al., 2017). The repeated measures can then be analyzed in wide format (with one row per individual and separate variables for each repeated measure) to allow for the clustering of repeated measures within an individual. This approach models the clustering of repeated measures in the imputation model by allowing for an unstructured pattern of correlations between them. However, this approach can only be used for repeated measures with follow-ups at fixed intervals of time within an individual, as in CATS.

Under this approach, because the repeated measures are imputed in wide format, the imputation model allows the relationship between the exposure and the outcome to be different at different time points, hence naturally incorporates an interaction between the time-varying exposure and time. Because this approach is congenial with an analysis model with an interaction between a time-varying covariate and time, we expect this approach to produce valid results for analysis model 1. In contrast, this approach does not implicitly accommodate interactions between the time-varying exposure and a time-fixed baseline variable (as in analysis model 2) or a nonlinear term (as in analysis model 3). To handle these analyses, additional terms can be incorporated into the imputation model using JAV or passively derived after imputation (i.e., the terms can be derived post-imputation or at each iteration, respectively, as described in the introduction), although we note that in this case the imputation model is not congenial with the substantive analysis model.

(ii) Single-level FCS with DI indicators for the higher level clusters and repeated measures imputed in wide format (FCS-IL-DI-wide)

An analogous approach to JM-IL-DI-wide is to use a single-level FCS approach with DI indicators representing the cluster membership of each individual, and imputing the repeated measures in the wide format. In contrast to JM, under the FCS approaches the imputations for the missing values in each variable in turn are drawn using an iterative algorithm that cycles through univariate imputation models (van Buuren et al., 2006). Similarly to JM-IL-DI-wide, DIs are included as predictors in these univariate models to model the correlation among individuals belonging to the same higher level cluster. Meanwhile the imputation model effectively also allows an unstructured correlation matrix between the repeated measures because, when imputing an incomplete repeated measure at one time point/wave, repeated measures at all the other waves are used as predictors.
This approach accommodates the interactions and the nonlinear terms considered in a similar manner to JM-1L-DI-wide.

(iii) **Two-level JM for the higher level clusters with repeated measures imputed in wide format (JM-2L-wide)**

First introduced by Schafer and Yucel (2002), the two-level JM approach is an extension of the single-level JM approach that imputes data using a joint MLMM (Schafer & Yucel, 2002). The JM-2L-wide approach consists of using a two-level JM to model the correlation among individuals within a higher level cluster using cluster-specific random effects (which are assumed to follow a normal distribution), with the repeated measures within individuals incorporated by imputing the data in wide format as in the previous approaches. This approach accommodates interactions and non-linear terms similarly to JM-1L-DI-wide and FCS-1L-DI-wide.

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

Van Buuren (2011) proposed an extension of FCS for imputing two-level data that uses a series of univariate two-level LMMs to impute the missing values, cycling through the incomplete variables one at a time (van Buuren, 2011). The FCS-2L-wide approach uses a univariate two-level LMM for each incomplete repeated measure with cluster-specific random effects to account for the correlation among individuals of the same higher level cluster, while repeated measures are imputed as distinct variables in wide format. Again this approach accommodates interactions and nonlinear terms similar to the above approaches.

(v) **Two-level substantive-model-compatible JM for the repeated measures with DI for the higher level clusters (SMC-JM-2L-DI)**

Goldstein et al. (2014) proposed an SMC-JM approach for imputing multilevel data, where the imputation model is defined as the product of the substantive model and the joint distribution of the covariates. With this approach, the imputations for incomplete variables are drawn simultaneously using an iterative Gibbs sampler algorithm with a Metropolis–Hastings step (Goldstein et al., 2014). Similarly to JM-1L-DI-wide and JM-2L-wide, the joint distribution is often assumed to be a MVN distribution, with categorical covariates imputed by assuming underlying latent continuous variables (Carpenter & Kenward, 2012; Goldstein et al., 2014). Current implementations of this approach can only handle up to two levels, that is, one type of clustering (Carpenter et al., 2011; Quartagno & Carpenter, 2016). Therefore in order to impute incomplete three-level data, when conducting the imputation step, the substantive model must be specified as a two-level (rather than three-level) LMM with a random effect modeling the repeated measures within an individual, and using DIs to represent the cluster membership. Here the two-level approach is used to handle the clustering of repeated measures within an individual as the substantive analysis model requires the time varying exposure to be in long format to model the linear trend with time that cannot be modeled with the data in the wide format. The correct (three-level) substantive model is fitted at the analysis stage to obtain the final parameter estimates. Under this approach, because the imputation model contains (a modified version of) the substantive model as a corresponding conditional, the interactions and the quadratic effects in our substantive analysis models are naturally incorporated in the imputation model.

(vi) **Two-level substantive-model-compatible sequential modeling for the repeated measures with DI for the higher level clusters (SMC-SM-2L-DI)**

Ibrahim et al. (2002) proposed an alternative SMC-SM approach where again the imputation model is decomposed into a product of the substantive model and the joint distribution of the covariates, but where the latter is further decomposed into a sequence of univariate conditional models for each incomplete covariate conditioning on the remaining variables (Ibrahim et al., 2002). Under this approach, each univariate conditional model can be specified as a generalized linear mixed model (GLMM) or a generalized linear model (GLM) according to the scale of measurement of the incomplete covariate. For example, a probit regression model can be used to impute binary variables and an ordered probit regression model for imputing ordinal variables (Lüdtke et al., 2019).
Current implementations of this approach available in R package “mdmb” can handle two-level incomplete data. Although it has been stated that the implementation may also be used to handle incomplete multilevel data with many levels, it is unclear how to do this with the information provided in the package manual. Therefore, similarly to SMC-SM-2L-DI, this method can be extended to handle incomplete three-level data by setting the substantive model to be a two-level GLMM in the imputation step, where the clustering of repeated measures within an individual is modeled via subject-specific effects (in long format), while the correlation among individuals of the same higher level cluster is modeled by DIs. Each incomplete covariate can then be modeled using an appropriate univariate conditional model. For incomplete time-varying (level 1) covariates, this would be a GLMM with subject-specific effects to model the within-individual correlation and DIs to model the higher level clustering, while with time-fixed (level 2) covariates a GLM with just the DIs to model the higher level clustering suffices. For modeling covariates measured at the higher level clusters (level 3), an appropriate GLM can be used. The correct substantive model is then fitted in the analysis step.

Under this approach the interactions and the quadratic effects are incorporated naturally in a similar manner to SMC-JM-2L-DI.

(vii) Three-level substantive-model-compatible JM (SMC-JM-3L)

Finally, Enders et al. (2019) have developed a three-level SMC approach, which is implemented in the Blimp software (Keller & Enders, 2019). Similarly to SMC-JM-2L-DI, the imputation model is decomposed into the product of the substantive model and the joint distribution of the covariates, but in this case the substantive model is defined as a three-level LMM. Within this approach, the correlation among individuals within the same higher level cluster is modeled using cluster-specific random effects and the correlation among repeated measures within individuals via subject-specific effects (in long format) (Enders et al., 2019). As with SMC-JM-2L-DI, the joint covariate distribution assumed is a MVN distribution, with categorical covariates handled by assuming underlying latent continuous variables. However, in contrast to SMC-JM-2L-DI, this approach imputes the incomplete covariates one at a time by factorizing the implied joint distribution of the covariates into the univariate conditional distribution of each incomplete covariate using Gibbs sampling with a final Metropolis–Hastings step. Incomplete time-varying (level 1) variables are imputed using a three-level random intercept model, regressing on all the other covariates. Time-fixed (level 2) variables are imputed using a two-level random intercept model regressing on all the other level 2 and level 3 variables, and the cluster means of the level 1 variables. Similarly, higher level cluster specific variables (level 3) are imputed using linear regression on all the level 3 covariates and the level 1 and level 2 cluster means. Of note, the cluster means can either be calculated as the arithmetic averages of the variable at the required level (known as the “manifest” approach) or can be modeled as normally distributed latent variables (known as the “latent” approach) (Keller, 2019).

Similarly to SMC-JM-2L-DI and SMC-SM-2L-DI, because the imputation model is specified in a way that includes the substantive model as conditional, the interactions and the quadratic effect in the substantive analysis models are naturally accommodated in the imputation model.

Table 2 provides a summary of all the approaches discussed above along with details of the software packages within which each approach is available.

Note: In this study, we do not consider any SMC-FCS approaches as multilevel extensions of this approach are not currently available and the single-level implementations cannot be carried out with the data in the wide format.

4 SIMULATION STUDY

4.1 Simulation of complete and missing data

We conducted a simulation study to compare the performance of the MI approaches detailed above in the context of the three-level analysis model outlined in Equations (1)–(3). Data were generated as described below to obtain simulated datasets with sample size 1200 students for each scenario. The number of simulations for each scenario was 1000, chosen to limit the Monte Carlo error for the coverage of nominal 95% confidence intervals to approximately 0.7% (Morris et al., 2019).
**TABLE 2** Summary of the imputation approaches for handling incomplete three-level data

| MI approach       | Paradigm | Type                     | Software* | How the two sources of clustering are handled | How the approach accommodate interactions/non-linear terms |
|-------------------|----------|--------------------------|-----------|-----------------------------------------------|----------------------------------------------------------|
| **JM-IL-DI-wide** | JM       | Standard (single-level)  | SAS, SPSS, Stata, Mplus, R | DI | Repeated measures imputed in wide format | 
| **FCS-1L-DI-wide** | FCS      | Standard (single-level)  | SAS, SPSS, Stata, Mplus, R, Blimp | DI | Repeated measures imputed in wide format | 
| **JM-2L-wide**    | JM       | Specialized for two levels | SAS, Mplus, Realcom-impute, Stat-JR, R | RE | Repeated measures imputed in wide format | 
| **FCS-2L-wide**   | FCS      | Specialized for two levels | Mplus, R, Blimp | RE | Repeated measures imputed in wide format | 
| **SMC-JM-2L-DI**  | JM       | Specialized for two-levels | R, Realcom-impute, Stat-JR | DI | RE | Through SMC-MI algorithm $^+$ | 
| **SMC-SM-2L-DI**  | SM       | Specialized for two-levels | R | DI | RE | Through SMC-MI algorithm $^+$ | 
| **SMC-JM-3L**     | JM       | Specialized for three-levels | Blimp | RE | RE | Through SMC-MI algorithm $^{++}$ | 

Abbreviations: DI, dummy indicators; FCS, fully conditional specification; JM, joint modelling; RE, random effects; SM, sequential modeling; SMC, substantive model compatible.

*The interactions or quadratic effects are accommodated as the imputation model specified under these approaches contains (a two-level version of) the substantive model as a corresponding conditional.

**Although the SM approach can be potentially used for multilevel data with an arbitrary number of levels that can be hierarchical or non-hierarchical, it is not clear how three-level data can be handled with the current information available about the package, see supplementary documentation from Lüdtke, Robitzsch and West (2019).

$^+$ The interactions and quadratic effects that are accommodated as the imputation model specified under this approach contains the substantive model as a corresponding conditional.
For each analysis model, we considered scenarios with two different cluster sizes: one scenario with 40 schools (clusters) each with 30 students, similar to CATS, and another with 10 clusters each with 120 students, to provide a scenario with a smaller number of higher level clusters. The variables were generated sequentially as described below for individual $j$ in cluster $i$. The values of the parameters indexing these distributions were determined by estimating the respective quantity from the CATS data and are given in Additional file 1: Table S2.

(i) Individuals were generated within school clusters under the two scenarios described above.
(ii) Child’s age at wave 1 ($age_{ij1}$) was generated from a uniform distribution, $U(7, 10)$.
(iii) Child’s sex ($sex_{ij}$) was generated by randomly assigning 4% of students to be female.
(iv) Child’s standardized SES value at wave 1 ($SES_{ij1}$) was generated from a standard normal distribution, $N(0, 1)$.
(v) The standardized NAPLAN scores at wave 1 ($NAPLAN_{ij1}$) were generated from a linear regression model conditional on child’s sex, child’s age at wave 1 and child’s SES value:

$$NAPLAN_{ij1} = \eta_0 + \eta_1 * sex_{ij} + \eta_2 * age_{ij1} + \eta_3 * SES_{ij1} + \psi_{ij}, \quad (4)$$

where $\psi_{ij}$ are independently and identically (iid) distributed as $\psi_{ij} \sim N(0, \sigma^2_\psi)$.
(vi) To generate the time-varying variables, each student record was expanded to include five repeated observations (from waves 2–6).
(vii) Child’s depression status at waves 2, 4, and 6 ($depression_{ij(k−1)}$) was generated using an LMM conditional on child’s age at wave 1, child’s sex, NAPLAN scores at wave 1, child’s SES value and wave:

$$depression_{ij(k)} = \delta_0 + \delta_1 * age_{ij1} + \delta_2 * sex_{ij} + \delta_3 * NAPLAN_{ij1} + \delta_4 * SES_{ij1} + \delta_5 * wave_{ijk} + u_{ij0} + u_{ij0} + \varphi_{ijk}, \quad (5)$$

where $\varphi_{ijk}, u_{ij0}, u_{ij0}$ are iid as $\varphi_{ij} \sim N(0, \sigma^2_\varphi), u_{ij0} \sim N(0, \sigma^2_u)$, and $u_{ij0} \sim N(0, \sigma^2_v)$ respectively.
(viii) The outcome, child’s standardized NAPLAN score at waves 3, 5, and 7 ($NAPLAN_{ij(k)}$), was generated using the relevant target analysis model as per Equations (1)–(3).
(ix) Finally, we generated child’s behavioral problems at waves 2, 4, and 6 ($SDQ_{ij(k)}$), which is not included in the analysis model but is associated with the exposure, using an LMM conditional on depression symptoms at waves 2, 4, and 6 and wave:

$$SDQ_{ij(k)} = \gamma_0 + \gamma_1 * depression_{ij(k)} + \gamma_2 * wave_{ijk} + v_{ij0} + v_{ij0} + \epsilon_{ij(k)} \quad (6)$$

where $\epsilon_{ij(k), v_{ij0}, v_{ij0}}$ are iid as $\epsilon_{ij(k)} \sim N(0, \sigma^2_\epsilon), v_{ij0} \sim N(0, \sigma^2_v)$, and $v_{ij0} \sim N(0, \sigma^2_v)$, respectively. This is an example of an auxiliary variable that can be included in the imputation model to improve its performance (Collins et al., 2001).

### 4.2 Generation of missing data

To simulate missingness, data were set to missing in depressive symptom scores at waves 2, 4, and 6 (the exposure of interest) and SES at wave 1. In our simulation study, for depression symptom scores, the proportions of missingness were set around 15%, 20%, and 30% at waves 2, 4, and 6, respectively, which were generated according to two MAR mechanisms, labeled MAR-CATS and MAR-inflated, by drawing from a logistic regression model dependent on the standardized NAPLAN scores at the subsequent wave and the SDQ measure at the concurrent wave:

$$\logit \{ P \left( R_{depression_{ij(k)} = 1} \right) \} = \zeta_0k + \zeta_1 * NAPLAN_{ij(k+1)} + \zeta_2 * SDQ_{ij(k)}, \quad (7)$$

where, $R_{depression_{ij(k)}}$ is an indicator variable that takes the value 0 if $depression_{ij(k)}$ is missing and 1 if $depression_{ij(k)}$ is observed.

For the MAR-CATS scenario, we set $\zeta_1 = 1.5$ and $\zeta_2 = 2$ that represent the associations between the probability of response and the predictors of response as observed in CATS. For the MAR-inflated scenario, the values of $\zeta_1$ and $\zeta_2$ were
doubled. The values of the intercepts $\zeta_{0k}$ were chosen by iteration so that the required proportions of missingness were achieved for each of the waves (2, 4, and 6). For simplicity, we set around 10% of the individuals to have missing SES values according to a missing completely at random (MCAR) mechanism using simple random sampling.

4.3 | MI methods and evaluation

For each of the 12 scenarios considered (3 substantive analysis models × 2 cluster sizes × 2 missingness mechanisms), we applied the seven MI approaches (JM-1L-DI-wide, FCS-1L-DI-wide, JM-2L-wide, FCS-2L-wide, SMC-JM-2L-DI, SMC-SM-2L-DI, and SMC-JM-3L), using ad hoc extensions to handle the interactions or nonlinear terms where necessary, to impute missing values in depressive symptom scores at waves 2, 4, and 6 and SES at wave 1 in each of the simulated data sets (as described below).

Specifically, for analysis model 1, which involves an interaction between the depressive symptom scores and time, no variations of the seven MI approaches were considered as all approaches are congenial with the analysis model.

For analysis model 2, which involves an interaction between the depressive symptom scores and SES, JM-1L-DI-wide and JM-2L-wide were applied using JAV to incorporate the interaction by including the interactions between depressive symptom scores at each wave (2, 4, and 6) and SES as distinct variables in the imputation model (denoted as JM-1L-DI-wide-JAV and JM-2L-wide-JAV). The JAV approach was used here as it has been shown in previous simulation studies to perform better than passive imputation post-imputation that is the only passive approach possible with JM (Seaman et al., 2012; Von Hippel, 2009). FCS-1L-DI-wide and FCS-2L-wide were implemented using passive imputation carried out within iterations using two variations of reverse imputation strategy (Grund et al., 2018; van Buuren, 2018):

(i) In the first approach (denoted FCS-1L-DI-wide-passive_c and FCS-2L-wide-passive_c), we included the single interaction between the NAPLAN score at the next wave and SES as a predictor in the imputation model for depressive symptom scores and the two-way interactions between the NAPLAN scores and depressive symptom scores at previous wave for all three waves as predictors when imputing SES. This method allows the association between the outcome and exposure at each wave to vary for different levels of SES and vice versa as implied by the substantive analysis model.

(ii) In the second approach (denoted FCS-1L-DI-wide-passive_all, FCS-2L-wide-passive_all), we included the two-way interactions between the NAPLAN scores at each of the three waves and SES as predictors in the imputation model for depressive symptoms at each wave and the two-way interactions between the NAPLAN scores and depressive symptom scores at previous wave for all three waves as predictors when imputing SES. Similar to the first approach, this method allows the association between the outcome and the exposure to vary for different levels of SES and vice versa, but allows even more flexibility. (Note: although it is possible to include all the two-way interactions between NAPLAN scores and depressive symptom across the waves as predictors when imputing LES allowing further flexibility, we opted to only include the ones implied by the analysis model for simplicity).

For analysis model 3, which involves a quadratic effect of the exposure, for JM-1L-DI-wide and JM-2L-wide we use a JAV approach for the quadratic term of the exposure, while for FCS-1L-DI-wide and FCS-2L-wide we use passive imputation of the quadratic term within iterations. For both analysis models 2 and 3, we also present results from the JM-1L-DI-wide and FCS-1L-DI-wide approaches where we do not incorporate the interaction or the quadratic term as a benchmark for comparison, resulting in a total of 11 and 9 approaches under analysis model 2 and 3 respectively.

Except for SMC-JM-3L in B1imp, all approaches were implemented in R version 3.6.1 (R Core Team, 2013). JM-1L-DI-wide, JM-2L-wide, and SMC-JM-2L-DI were implemented in the R package “jomo” using the functions jomocon, jomo1rancon, and jomo.lmer respectively, while FCS-1L-DI-wide and FCS-2L-wide were implemented in the R package “mice” using the functions norm and 2l.pan, respectively (Quartagno & Carpenter, 2016; van Buuren & Groothuis-Oudshoorn, 2010). Although there are several alternative functions in R for specifying a two-level FCS imputation model, we chose the above because they are the most established. The sequential modeling approach, SMC-SM-2L-DI, was implemented using the function frm_fb in the R package “ndmb” (Robitzsch & Lüdtke, 2018). For SMC-JM-3L, we used the B1imp version 1.0.6 with the default “latent” specification (Keller & Enders, 2019).

In addition to all the variables in the analysis model, the imputation model for each of the MI approaches also included child behavior problems (measured by SDQ) at waves 2, 4, and 6 as auxiliary variables. For each simulated dataset, 20
imputations were generated for each of the MI approaches (Carpenter & Kenward, 2012). After examining trace plots, the JM and SMC-SM approaches in R were run with a burn-in of 1000 iterations and 100 between-imputation iterations, while the SMC-JM approach in R were run with a burn-in of 500 iterations and 10 between-imputation iterations. The FCS approaches were run with a burn-in of 10 iterations. The SMC-JM approach in Blimp was run with a burn-in of 2500 iterations with 10 thinning iterations, as with 2500 iterations the highest (worst) potential scale reduction (PSR) factor value across all parameters was generally less than 1.10 (Enders et al., 2018). These values were further confirmed by examining trace plots.

The substantive analysis models were fitted to each of the imputed data sets and the resulting parameter estimates pooled using Rubin’s rules (Grund et al., 2017; Rubin, 1987). For each analysis model, the parameters of interest were the regression coefficients corresponding to the main effect of the exposure and the interaction or quadratic term, and the variance components at levels 1, 2 and 3. The estimates of these parameters obtained from the MI approaches were compared to the true values that were used to simulate the data. In order to compare the performance of the various approaches for estimating the regression coefficients of interest we calculated, across 1000 replications, the mean value of the estimate, the bias (the average difference between the true value and the estimates), the empirical standard error (the average standard deviation of the estimates), the model-based standard error (the average of the estimated standard errors 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) (Burton et al., 2006). For the variance component estimates, we report the bias and empirical standard error. We also report the percentage bias defined as the bias relative to the true value.

An overview of the simulation design highlighting the aims, data-generating mechanisms, estimands, methods, and performance measures (“ADEMP” structure) (Morris et al., 2019), along with the table and figure numbers of the results from different simulation conditions are provided in Table S1.

4.4 | Simulation results

As the comparative performance of the MI approaches was quite similar for the MAR-CATS and MAR-inflated scenarios, we largely focus on the results from the MAR-CATS scenario highlighting contrasts where they exist.

(i) Analysis model 1: Interaction between the time-varying exposure and time

The sampling distribution of the estimated bias of the regression coefficients of interest for analysis model 1 (the main effect $\beta_1$ and the interaction effect $\beta_3$) across the 1000 replications for each MI approach, for the two different higher level cluster settings, is displayed in Figure 1a. Although all the MI approaches produced approximately unbiased estimates of these parameters, slightly higher biases were observed for SMC-SM-2L-DI and SMC-JM-3L.

All the MI approaches also resulted in appropriate nominal coverage and comparable empirical and model-based standard errors for both of the regression coefficients across the two simulation scenarios (Figure 1b and Table S3). However, with a larger number of higher level clusters SMC-JM-2L-DI resulted in slightly higher empirical standard errors compared with model-based standard errors.

Figure 1c shows the sampling distribution of the estimated biases for the variance components at levels 1, 2 and 3 across the two simulation scenarios. All approaches resulted in similar negligible bias (< 10% relative bias) for the three variance components for both scenarios with slightly larger biases for the level 3 variance estimates when there was a smaller number of higher level clusters.

(ii) Analysis model 2: Interaction between the time-varying exposure and a time-fixed baseline variable

For analysis model 2, the estimates of the main effect were approximately unbiased across all MI approaches, but there were substantial differences in the bias for estimating the interaction effect (Figure 2a). As expected, JM-IL-DI-wide and FCS-IL-DI-wide (which do not incorporate the interaction in the imputation model) resulted in the largest bias for the interaction term, attenuating it toward zero (Table S7). Although including the interaction term as a distinct variable within the JM approaches (JM-IL-DI-wide-JAV and JM-2L-wide-JAV) and using passive imputation within the FCS approaches (FCS-IL-DI-wide-passive_c, FCS-IL-DI-wide-passive_all, FCS-2L-wide-passive_c and FCS-2L-wide-passive_all)
FIGURE 1  (a) Distribution of the bias in the estimated regression coefficient for the main effect ($\hat{\beta}_1$, true value $=-0.07$) and the interaction effect ($\hat{\beta}_3$, true value $=0.013$); (b) Empirical standard error (filled circles with error bars showing $\pm 1.96 \times$ Monte Carlo standard errors) and average model-based standard error (hollow circles); (c) Distribution of the bias in the estimated variance components at level 1–3 in analysis model 1 across the 1000 simulated datasets for the seven multiple imputation (MI) approaches under two scenarios for number of higher level clusters (40 school clusters and 10 school clusters) when data are missing at random with relationships based on the CATS data (MAR-CATS). The lower and upper margins of the boxes in (a) and (b) represent the 25th (Q1) and the 75th (Q3) percentiles of the distribution, respectively. The whiskers extend to Q1-$1.5*(Q3-Q1)$ at the bottom and Q3+$1.5*(Q3-Q1)$ at the top. The following abbreviations are used to denote different MI methods, for example, DI: dummy indicators, FCS: fully conditional specification, JM: joint modeling, SM: sequential modelling, SMC: substantive model compatible

resulted in reduced bias compared to the naïve methods (JM-1L-DI-wide and FCS-1L-DI-wide), these approaches produced larger bias than the SMC MI approaches (SMC-JM-2L-DI, SMC-SM-2L-DI and SMC-JM-3L). The bias in the interaction effect estimates under the FCS approaches that include all the interactions as predictors when imputing the incomplete depressive symptom scores (FCS-1L-DI-wide-passive_all and FCS-2L-wide-passive_all) were smaller than the FCS approaches that include the single interaction as implied under the analysis model (FCS-1L-DI-wide-passive_c and FCS-2L-wide-passive_c).
FIGURE 2  (a) Distribution of the bias in the estimated regression coefficient for the main effect ($\hat{\beta}_1$, true value = $-0.024$) and the interaction effect ($\hat{\beta}_3$, true value = $0.023$); (b) Empirical standard error (filled circles with error bars showing ±1.96x Monte Carlo standard errors) and average model-based standard error (hollow circles); (c) Distribution of the bias in the estimated variance components at level 1–3 in analysis model 2 across the 1000 simulated datasets for the 11 multiple imputation (MI) approaches under two scenarios for number of higher-level clusters (40 school clusters and 10 school clusters) when data are missing at random with relationships based on the CATS data (MAR-CATS). The lower and upper margins of the boxes in (a) and (b) represent the 25th (Q1) and the 75th (Q3) percentiles of the distribution, respectively. The whiskers extend to Q1-1.5*(Q3-Q1) at the bottom and Q3 +1.5*(Q3-Q1) at the top. The following abbreviations are used to denote different MI methods, for example, DI: dummy indicators, FCS: fully conditional specification, JM: joint modelling, SM: sequential modeling, SMC: substantive model compatible.
Differences between the model-based and empirical standard errors for the main effect were small for all MI approaches, with somewhat larger discrepancies when the number of higher-level clusters was large under the MAR-inflated scenario (Figure S5). Although JM-IL-DI-wide-JAV, JM-2L-wide-JAV, SMC-SM-2L-DI and SMC-JM-3L provided model-based standard errors that were largely consistent with the empirical standard errors for the interaction effect, JM-IL-DI-wide-JAV and JM-2L-wide-JAV resulted in upward-biased estimates of the model-based standard errors. Somewhat larger discrepancies between the model-based and empirical standard errors were observed for the interaction effect for all of the other MI approaches across the simulation scenarios considered (Figure Sb, Table S7 and S9).

All of the MI approaches resulted in negligible bias (< 10% relative bias) for the variance components at levels 1, 2 and 3 across the different simulation scenarios, albeit slightly larger for the level 3 and level 2 variance components when there were fewer higher level clusters (Figure 2c, Table S8).

(iii) Analysis model 3: Quadratic term in the exposure

All of the MI approaches except for SMC-JM-2L-DI, SMC-SM-2L-DI, and SMC-JM-3L resulted in biased estimates of the regression coefficients for the main effect and the quadratic term in analysis model 3, with substantial underestimation of the quadratic effect (Figure 3a and Table S10). Discrepancies between the model-based and empirical standard errors for the quadratic effect were observed for JM-IL-DI-wide, FCS-IL-DI-wide, FCS-IL-DI-wide-passive, and FCS-2L-wide-passive (Figure 3b). However, these approaches retained approximate nominal coverage for the quadratic term under all scenarios. JM approaches using JAV to handle the quadratic effect (JM-IL-DI-wide-JAV and JM-2L-wide-JAV) resulted in upward-biased model-based standard errors for both coefficient estimates compared to other MI approaches. Under-coverage of the quadratic effect was also observed for these two approaches that was more pronounced under MAR-inflated scenario (Table S11).

Figure 3c shows the estimated bias for the variance components at levels 1, 2 and 3 across different simulation scenarios. All approaches resulted in negligible bias (< 10% relative bias) for the variance components across the different simulation scenarios, with slightly larger bias for the levels 3 and 2 variance estimates when there was a smaller number of higher level clusters (Tables S12 and S14).

5 | CATS CASE STUDY ILLUSTRATION

We applied the same MI approaches (7, 11, or 9 depending on the analysis model) as used in the simulations to the three target analysis models in the CATS data. Twenty imputations were generated under each of the MI approaches. Similarly to the simulation study, the imputation model for each of the MI approaches included all the variables in the analysis model and also the child behavior problems measures (reported by SDQ) at waves 2, 4, and 6 as auxiliary variables. However, in the CATS, missing values occurred in the outcome (NAPLAN scores at waves 3, 5, and 7), the exposure (depressive symptom scores at waves 2, 4, and 6), and a time-fixed baseline variable (NAPLAN scores at wave 1). The auxiliary variable (SDQ at waves 2, 4, and 6) was also incomplete with values missing for 29% (363/1239) of individuals at wave 2, 35% (436/1239) at wave 4, and 24% (297/1239) at wave 6. We do not include further auxiliary variables specifically for imputing the outcome. The SDQ variable is also correlated with the outcome to some degree and hence would provide some auxiliary information. Figures 4-6 illustrate the estimated regression coefficients and the 95% confidence interval for the main effect and the interaction/quadratic effect for analysis models 1–3, respectively, under the different imputation approaches, while Tables S15–S17 show the estimated regression coefficient along with the standard errors and the variance component estimates.

(i) Analysis model 1: Interaction between the time-varying exposure and time

As shown in Figure 4, the estimates for the main effect and the corresponding standard errors were very similar for all the MI approaches except for JM-IL-DI-wide and SMC-SM-2L-DI, which resulted in smaller estimates. Estimates of the interaction effect and standard errors were similar across all approaches, while estimates of the level 3 variance component were smaller for SMC-JM-2L-DI and SMC-SM-2L-DI, with comparatively larger level 2 variance components for SMC-JM-2L-DI (Table S10).

(ii) Analysis model 2: Interaction between the time-varying exposure and a time-fixed baseline variable
Figure 3 (a) Distribution of the bias in the estimated regression coefficient for the main effect ($\beta_1$, true value $= -0.024$) and the quadratic effect ($\beta_3$, true value $= -0.009$); (b) Empirical standard error (filled circles with error bars showing $\pm 1.96 \times$ Monte Carlo standard errors) and average model-based standard error (hollow circles); (c) Distribution of the bias in the estimated variance components at level 1–3 in analysis model 3 across the 1000 simulated datasets for the nine multiple imputation (MI) approaches under two scenarios for number of higher level clusters (40 school clusters and 10 school clusters) when data are missing at random with relationships based on the CATS data (MAR-CATS). The lower and upper margins of the boxes in (a) and (b) represent the 25th (Q1) and the 75th (Q3) percentiles of the distribution, respectively. The whiskers extend to Q1-1.5*(Q3- Q1) at the bottom and Q3 +1.5*(Q3- Q1) at the top. The following abbreviations are used to denote different MI methods, for example, DI: dummy indicators, FCS: fully conditional specification, JM: joint modeling, SM: sequential modeling, SMC: substantive model compatible.

Figure 5 and Table S11 show the results from the 11 MI approaches applied to analysis model 2 in the CATS data. The main effect coefficient estimates and their standard errors were very similar irrespective of the analysis method. Although the estimates for the interaction effect were similar across most approaches, larger estimates were observed for JM-1L-DI-wide-JAV and JM-2L-wide-JAV. Similar to analysis model 1, variance component estimates were generally similar although there were smaller estimates for the level 3 variance using SMC-JM-2L-DI and SMC-SM-2L-DI with slightly larger estimates for the level 2 variance component for SMC-SM-2L-DI (Table S11).
(iii) Analysis model 3: Nonlinear relationship with the exposure

The results from the MI approaches applied to analysis model 3 in the CATS data are shown in Figure 6 and Table S12. In comparison to the other MI approaches, the regression coefficient estimates for the main effect were larger while the quadratic effect estimates were smaller for JM-IL-DI-wide, FCS-IL-DI-wide, and FCS-IL-DI-wide-passive. Similarly to the analysis of models 1 and 2, the variance component estimates were generally similar across the approaches, with smaller estimates for the level 3 variance components with SMC-JM-2L-DI and SMC-SM-2L-DI and comparatively larger estimates for the level 2 variance component for SMC-SM-2L-DI (Table S12).
6 | DISCUSSION

Although adaptations to single- and two-level MI approaches, using DIs and imputing in wide format, and three-level MI approaches can all be used to accommodate the sources of correlation in three-level data in the imputation process, when the substantive analysis model includes interactions or quadratic effects involving incomplete covariates, these too need to be incorporated in the imputation model. Although SMC MI has shown great promise in accommodating such terms, there are limited SMC MI implementations that explicitly handle three-level data. In this study, we evaluated the performance of several adaptations to currently available MI approaches and the single implementation of SMC MI that was designed to accommodate three-level data, for handling incomplete three-level data in the context of three commonly used LMMs in the analysis of longitudinal data that include an interaction or a nonlinear term via simulations and a real data example based on the CATS.

When the analysis model included an interaction between the exposure and time, all of the MI approaches resulted in approximately unbiased estimates of the main effect and the interaction effect, with appropriate coverage, across the different simulation scenarios considered in the simulation study. However, when the analysis models involved an interaction between the exposure and a baseline confounder, or a quadratic effect of the exposure, the approaches that used ad hoc extensions of single- and two-level models resulted in biased estimates of the interaction and the nonlinear effects. In contrast, the two-level SMC approaches extended with DIs to handle the second level of clustering, SMC-JM-2L-DI and SMC-SM-2L-DI, showed similar performance to the three-level SMC approach, SMC-JM-3L, all of which resulted in approximately unbiased estimates of the interaction or nonlinear effects and the variance components. However, in the CATS application the former two approaches resulted in comparatively smaller level 3 variance components than
the latter. These results suggest that substantive model compatibility is crucial for ensuring appropriate imputations as shown in previous literature in different contexts (Bartlett et al., 2015; Enders et al., 2019; Erler et al., 2016; Goldstein et al., 2014; Kim et al., 2015).

Although there have been recent developments in SMC MI approaches that have shown great promise for accommodating interactions and/or nonlinear terms appropriately in the MI literature, implementations of these approaches for multilevel data are limited. In fact, commonly used statistical packages Stata and SPSS do not have SMC-MI approaches, or indeed any MI approaches, which use multilevel imputation models. Therefore, when the substantive analysis model involves an interaction with time; for users of these packages, the only option is to use the single-level MI approaches with pragmatic adaptations, namely JM-IL-DI-wide and FCS-IL-DI-wide. However, both of these approaches can only be used when repeated measures are recorded at fixed intervals of time. In addition, although we observed approximately unbiased results from these approaches, previous simulation studies have shown the DI approach can result in inflated standard errors and biased variance components estimates (and therefore ICCs) particularly when the ICC is low with a high percentage of missing values and there are small cluster sizes (Drechsler, 2015; Lüdtke et al., 2017). Therefore, these approaches should be used with caution. Both these approaches can also be infeasible with a large number of clusters.

Our results confirm, as expected theoretically and from past simulations (Enders et al., 2019; Keller, 2019), that the three-level SMC approach, SMC-JM-3L, is the most appropriate approach for handling incomplete three-level data where the analysis model includes an interaction among incomplete covariates or quadratic effects. However, this approach is only available in the stand-alone software B1imp (Keller & Enders, 2019). Although the two-level SMC MI approaches with DIS used to handle the higher level clustering, SMC-JM-2L-DI and SMC-SM-2L-DI, may be potentially useful alternatives, as made evident by the CATS application in our study and previous simulation studies, the DI approach can be problematic when the ICC is low and there is a high percentage of missing values or if there are small cluster sizes (Drechsler, 2015; Lüdtke et al., 2017). We also note that with these two approaches, a modified version of the analysis model, must be used in the imputation procedure and the actual substantive analysis model needs to be fitted to the imputed data following imputation. This is in contrast to when these approaches are used to impute two-level data, where the parameters of interest in the substantive analysis model are estimated within the imputation process itself (Erler et al., 2016). The SMC-SM-2L-DI approach is somewhat more flexible than SMC-JM approaches because each incomplete covariate is modeled separately, which means that nonlinear associations among covariates can also be accommodated in the imputation model. This comes at a cost, as this approach requires specification of separate models for each incomplete covariate and ordering of these models that requires more consideration than the specification of the imputation model under SMC-JM approaches (Erler et al., 2019). One strategy suggested is to order the incomplete variables by their type and start by conditioning the categorical variables on the continuous variables. Other strategies include ordering the variables so that the missing pattern is close to monotone or according to the percentage of missing data, that is, starting with those with the least percentage of missing values (Bartlett et al., 2015; Erler et al., 2019; Lüdtke et al., 2019). Although, theoretically, the order in which the conditional models for the covariates are specified may result in different joint distributions, the approach has been shown to be robust under different orderings as long as the conditional models for each covariate “fit the data well enough” (Erler et al., 2016).

In addition to the differences discussed previously, there are also some differences between the two-level JM approach by Schafer and Yucel (2002), as implemented in R, and the two-level FCS implementation in R as originally proposed by van Buuren (2011). The multilevel JM approach allows associations between incomplete variables to vary at different levels, whereas the FCS approach does not allow this (Enders et al., 2016). The two approaches can, however, be made equivalent by including the cluster means of the imputed lower level variables in the FCS imputation model. This difference can be important in the context of a multilevel substantive analyses that assumes different associations between variables at different levels. Given the substantive analysis model considered in the current article does not assume such relations, these differences were not relevant here. Further discussion of the differences between the JM and FCS approaches in the multilevel context can be found elsewhere (Carpenter & Kenward, 2012; Enders et al., 2016; Mistler & Enders, 2017).

To our knowledge, there are no previous studies that compare all of the available MI approaches for accommodating interaction and/or nonlinear terms in a three-level data setting. However, the results presented in our study are consistent with previously published studies in the single- and two-level settings. Similarly to our findings, several simulation studies in the single-level context have shown that the JAV approach results in biased estimates of the interaction effect and some undercoverage when data are MAR (Bartlett et al., 2015; Seaman et al., 2012). Our results are also consistent with those of Grund (2018), who showed that passive imputation within iterations using the reverse imputation strategy
to accommodate interactions in a two-level model resulted in biased estimates of the interaction effect in the presence of incomplete covariates (Grund et al., 2018). In contrast, however, Tilling et al. (2016) reported approximately unbiased results using a reverse imputation strategy with the JAV approach, their simulations involving an interaction between two binary covariates (Tilling et al., 2016). Consistent with our results, several studies in the single-level context have also shown the SMC MI approaches perform better than MI approaches with ad hoc extensions such as JAV for accommodating interactions or nonlinear effects (Kim et al., 2015; Lüdtke et al., 2019; Zhang & Wang, 2017). Finally, Enders et al. (2019) and Keller (2019) presented a number of simulations evaluating the performance of the SMC-JM-3L implementation in Blimp against the JAV approach in a wide range of single-level and multilevel regression models with interaction and nonlinear effects, which showed that SMC-JM-3L generally resulted in approximately unbiased estimates of the model parameters (Enders et al., 2019; Keller, 2019).

Designing our simulation study on a real study allowed us to incorporate complex yet realistic associations, hence we believe our results reflect what could be expected in practice in a similar setting. However, we recognize that the simulation conditions we examined are limited in scope. There are a number of factors at play when analyzing multilevel data such as the cluster sizes, number of clusters, intra-class correlations, effect sizes, and the variability of the random effects; and the performance of the methods evaluated in our study may vary with these factors. Therefore, caution is required in generalizing these results to conditions outside those evaluated in our study. Our study is restricted to substantive analyses where the interactions involve incomplete continuous predictors. When the interaction term involves different types of variables, we expect the SMC-JM-3L to perform well based on previous simulation studies (Keller, 2019), but the performance of other MI approaches may vary, particularly those that use the DI extension. In the special case, when one of the variables involved in the interaction is fully observed and categorical, an alternative approach that can be used, provided that there are not too many categories, is to split the data into strata as defined by this variable and carry out imputation separately within each stratum (Tilling et al., 2016). Given the limited number of three-level SMC MI implementations, a useful extension of our study would be to evaluate the performance of all the available imputation approaches for incomplete categorical variables in three-level data, as well as the 3-level approach in the R command “mmdab” that is apparently available but lacks implementation details. Our simulations were also limited to a random-intercept model. Previous research suggests that SMC-MI methods perform well in accommodating random slopes in multilevel models but we expect the performance of the adaptations in this context to be different (Enders et al., 2019; Huque et al., 2019). However, this was beyond the scope of our paper and still an area for future research. Finally, our simulations only considered MAR missingness mechanisms while in practice the data may be missing not at random (MNAR). Although the SMC-SM approach may be adapted to handle MNAR data (Lüdtke et al., 2019), all of the MI approaches considered in our simulations are only guaranteed to produce unbiased estimates under MAR (Moreno-Betancur et al., 2018). Therefore, examining the performance of these methods under MNAR mechanisms may also be an avenue for future research.

In our simulation design, data were generated by repeatedly drawing from the target analysis models of interest. Therefore, another limitation of this work is that we restrict our attention to the scenario where the substantive analysis model is correctly specified, and hence the results presented would only be valid under this assumption. Although this is a very strong assumption that may not be realistic in practice, it was necessary to address the aim of our research that is about evaluating the performance of the various MI methods without the results being influenced by misspecification of the analysis model. This approach obviously favors an SMC approach. A more neutral comparison of the approaches may be conducted using neutral comparison studies, as defined by Boulesteix et al. (2017), which are designed ensuring that no one method is favored. However, in practice such studies can be time consuming and difficult to perform (Boulesteix et al., 2017). As future work, examining the effect of substantive model misspecifications on the SMC approaches would be beneficial. In this paper, we have focused on the method of MI to handle the missing data. An alternative approach would be to use a direct Bayesian analysis to obtain the posterior distribution of the target parameters of interest. Such analyses can be performed using dedicated Bayesian analysis packages such as OpenBUGS, JAGS, and R (STAN) to name a few.

In conclusion, in this study we have shown that the single- and two-level MI approaches, or two-level SMC-MI approaches extended with DIs and/or imputing repeated measures in wide format perform as well as the three-level SMC-MI approaches in accommodating interactions between time-varying exposures and time in the substantive analysis model. Hence in such a context, practitioners may use any of these approaches. However, approaches that use the DI extension should be used with caution as they can be problematic in certain scenarios. We recommend SMC three-level MI approaches be used for handling incomplete three-level data when the substantive analysis includes an interaction between the time-varying exposure and an incomplete time-fixed confounder, or a quadratic effect of the exposure.
ACKNOWLEDGMENTS
This work was supported by funding from the National Health and Medical Research Council: Career Development Fellowship ID#1127984 (KJL) and project grant, ID#APP1166023. Research at the Murdoch Children’s Research Institute is supported by the Victorian Government’s Operational Infrastructure Support Program. MMB is 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.

CONFLICT OF INTEREST
The authors have declared no conflict of interest.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
For the simulation study, all data were simulated, which did not require ethical approval or consent from participants. The case study example used in this study was based on the Childhood to Adolescence Transition Study (CATS) that has been provided ethical clearance by the Royal Children’s Hospital Human Research Ethics Committee (#31089). Permission was granted from the Victorian Department of Education and Training and the Catholic Education Office Melbourne to recruit through their schools. Children were recruited through active, written, informed parental consent.

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 for simulating the data that support the findings of this study are openly available in a public GitHub repository at https://github.com/rushwije/MI_three-level.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID
Rushani Wijesuriya https://orcid.org/0000-0003-1023-4065
John B. Carlin https://orcid.org/0000-0002-2694-9463
Anurika P. De Silva https://orcid.org/0000-0003-0541-3202

REFERENCES
Allison, P. (2002). Missing data. Sage.
Angold, A., Costello, E. J., & Worthman, C. M. J. P. M. (1998). Puberty and depression: The roles of age, pubertal status and pubertal timing. Psychological Medicine, 28(1), 51–61. https://doi.org/10.1017/S003329179700593X
Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research, 24(4), 462–487. https://doi.org/10.1177/0962280214521348
Black, A. C., Harel, O., & Betsy McCoach, D. (2011). Missing data techniques for multilevel data: Implications of model misspecification. Journal of Applied Statistics, 38(9), 1845–1865. https://doi.org/10.1080/02664763.2010.529882
Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. J. B. M. R. M. (2017). Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. BiMC Medical Research Methodology [Electronic Resource], 17(1), 1–12. https://doi.org/10.1186/s12874-017-0417-2
Boulesteix, A. L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. J. B. J. B. Z. (2017). On the necessity and design of studies comparing statistical methods. Biometrical Journal, 60(1), 216–218. https://doi.org/10.1002/bimj.201700129
Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. Statistics in Medicine, 25(24), 4279–4292. https://doi.org/10.1002/sim.2673
Carpenter, J., & Kenward, M. (2012). Multiple imputation and its application: John Wiley & Sons.
Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. Journal of Statistical Software, 43(5), 1–14. https://doi.org/10.18637/jss.v043.i05
Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4), 330. https://doi.org/10.1037/1082-989X.6.4.330

Drechsler, J. (2015). Multiple imputation of multilevel missing data—Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1), 69–95. https://doi.org/10.3102/1076998614563393

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, 25(1), 88–112. https://doi.org/10.1037/met0000228

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2), 298–317. https://doi.org/10.1037/met0000148

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2). https://doi.org/10.1037/met0000063

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, 28(2), 555–568. https://doi.org/10.1177/0962280217730851

Erler, N. S., Rizopoulos, D., Rosmalen, J. V., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, 35(17), 2955–2974. https://doi.org/10.1002/sim.6944

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3), 173–197. https://doi.org/10.1177/1471082X0800900301

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(2), 553–564. https://doi.org/10.1111/rssa.12022

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40(11), 1337–1345.

Grund, S., Lüdtke, O., Robitzsch, A. (2018). Multiple Imputation of Missing Data for Multilevel Models. *Organizational Research Methods*, 21(1), 111–149. https://doi.org/10.1177/1094428117703686

Grund, S., Robitzsch, A., & Luedtke, O. (2017). MITML: Tools for multiple imputation in multilevel modeling. Version 0.4-3. https://doi.org/10.1177/1094428117703686

Hayes, T. (2019). Flexible, free software for multilevel multiple imputation: a review of Blimp and jomo. *Journal of Educational and Behavioral Statistics*, 44(5), 625–641. https://doi.org/10.1080/1076998619858624

Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B., Lee, K. J. (2020). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal*, 62(2), 444–466. https://doi.org/10.1002/bimj.201900051

Ibrahim, J. G., Chen, M. H., & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, 30(1), 55–78. https://doi.org/10.2307/3315865

Keller, B., & Enders, C. (2019). Blimp user’s guide (version 2.0). Los Angeles, CA.

Keller, B. T. (2019). *Model-based imputation for multilevel interaction effects*. UCLA.

Kim, S., Sugar, C. A., & Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine*, 34(11), 1876–1888. https://doi.org/10.1002/sim.6435

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1), 141. https://doi.org/10.1037/met0000996

Lüdtke, O., Robitzsch, A., West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, 25(2), 157–181. https://doi.org/10.1037/met0000233

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

Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, 42(4), 432–466. https://doi.org/10.1037/jeb0000169

Moreno-Betancur, M., Lee, K. J., Leacy, F. P., White, I. R., Simpson, J. A., & Carlin, J. B. (2018). Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *American Journal of Epidemiology*, 187(12), 2705–2715. https://doi.org/10.1093/aje/kwy173

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mundy, L. K., Canterford, L., Tucker, D., Bayer, J., Romaniuk, H., Sawyer, S., Lietz, P., Redmond, G., Proimos, J., Allen, N., & Patton, G. (2017). Academic performance in primary school children with common emotional and behavioral problems. *Journal of School Health*, 87(8), 593–601. https://doi.org/10.1111/josh.12531

Mundy, L. K., Simmons, J. G., Allen, N. B., Viner, R. M., Bayer, J. K., Olds, T., Williams, J., Olsson, C., Romaniuk, H., Mensah, F., Sawyer, S. M., Degenhardt, L., Alati, R., Wake, M., Jacka, F., & Patton, G. C. (2013). Study protocol: The childhood to adolescence transition study (CATS). *BMC Pediatrics*, 13(1), 160. https://doi.org/10.1186/1471-2431-13-160

Quartagno, M., & Carpenter, J. (2016). jomo: a package for joint modelling multiple imputation. R package version 2.2-0.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85–96.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, 15(1), 30. https://doi.org/10.1186/s12874-015-0022-1
Robitzsch, A., & Lüdtke, O. (2018). mdmb: Model based treatment of missing data (R package Version 1.0-18).
https://doi.org/10.1777/1536867x0400400301

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. Paper presented at the Proceedings of the survey research methods section of the American Statistical Association.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse. Wiley.

Schafer, J. L. (1997). Analysis of incomplete multivariate data: Chapman and Hall/CRC.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1), 19–35. https://doi.org/10.1111/1467-9574.00218

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMJ Medical Research Methodology*, 12(1), 46. https://doi.org/10.3758/s00211-017-0951-1

Speidel, M., Drechsler, J., & Sakshaug, J. W. (2018). Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. *Behavior Research Methods*, 50(5), 1824–1840. https://doi.org/10.3758/s13428-017-0951-1

Tilling, K., Williamson, E. J., Spratt, M., Sterne, J. A., & Carpenter, J. R. (2016). Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *Journal of Clinical Epidemiology*, 80, 107–115. https://doi.org/10.1016/j.jclinepi.2016.07.004

Van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Routledge.

Van Buuren, S. (2018). *Flexible imputation of missing data*: Chapman and Hall/CRC.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. https://doi.org/10.1080/10629360600810434

Vink, G., & van Buuren, S. (2013). Multiple imputation of squared terms. *Sociological Methods & Research*, 42(4), 598–607.

Von Hippel, P. T. (2009). 8. How to Impute Interactions, Squares, and other Transformed Variables. *Sociological Methodology*, 39(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4), 377–399. https://doi.org/10.1002/sim.4067

Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., & Lee, K. J. (2020). Evaluation of approaches for multiple imputation of three-level data. *BMC Medical Research Methodology*, 20(1), 1–15. https://doi.org/10.1186/s12874-020-01079-8

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, 22(4), 649–666. https://doi.org/10.1037/met0000104

**Supporting Information**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., De Silva, A. P., & Lee, K. J. (2022). Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data. *Biometrical Journal*, 64, 1404–1425. https://doi.org/10.1002/bimj.202000343