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

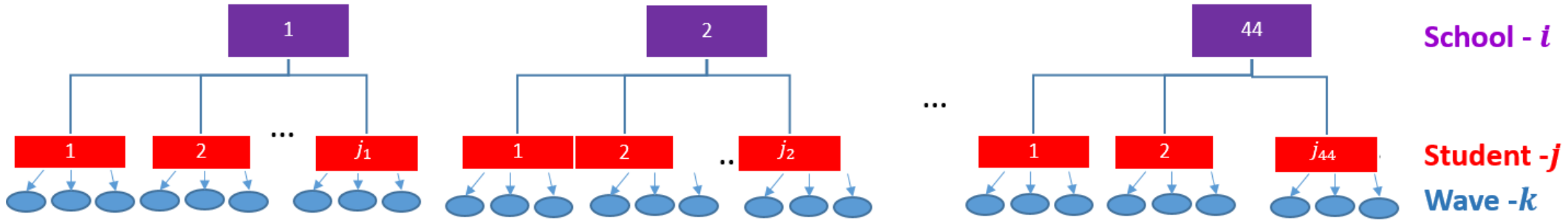
Rushani Wijesuriya

Prof Katherine Lee, Dr. Margarita Moreno-Betancur, Prof John Carlin and Dr. Anurika De Silva

26th of August 2020

Background

Three-level data structures are common in clinical and population health research, and Childhood to Adolescence Transition Study (CATS) is one such study where repeated measures (level 1) of students (level 2) are nested within schools (level 3)



In longitudinal studies like CATS, missing data is a major problem

Multiple imputation (MI) is a popular approach for handling missing data(1)

A key consideration in MI is that in order to generate valid inferences, the imputation model needs to preserve all the features of the analysis model such as non-linear relationships, interactions and multilevel features(2)(3)

(1) 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 med res methodology*, 15, 30.
(2) Meng, X.-L. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538-558
(3) 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, 462-487

Background

Accommodating the three-level structure and non-linear or interactions in the imputation model

Accommodating the three-level structure[4]

Extend single-level MI approaches	Extend two-level MI approaches	Extend two-level MI approaches	Use three-level MI approaches/Mixed model based MI (repeated measures imputed in long format)
• School clusters :Dummy indicators (DI)* • Repeated measures: imputed in wide format	• School clusters : Mixed model based MI • Repeated measures: imputed in wide format	• School clusters: Dummy indicators (DI)* • Repeated measures: Mixed model based MI (imputed in long format)	

Accommodating interactions or non-linear terms

As repeated measures are in wide format (unless the interaction is with time) ad-hoc extensions will need to be used: • Impute these terms as Just another variable (JAV) • passively impute these terms after imputation or at each iteration		As the repeated measures are in long format substantive model compatible (SMC) MI can be used	
JM-1L-DI-wide ***FCS-1L-DI-wide***	***JM-2L-wide*** ***FCS-2L-wide***	***SMC-JM-2L-DI*** ***SMC-SM-2L-DI***[5]	***SMC-JM-3L***[6]

*DI extension should be used with caution as it has been shown to produce biased parameter estimates in certain scenarios in some MI literature

*FCS: fully conditional specification, JM: joint modelling, SM: sequential modelling,

(4) Wijesuriya, R. *et al.* Evaluation of approaches for multiple imputation of three-level data. *BMC Med Res Methodology* **20,** 207 (2020).

(5) Lüdtke, O., Robitzsch, A. & West, S. G. 2019. Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological methods*

(6) Enders, C. K., Du, H. & Keller, B. T. 2019. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological methods*

Aim

- The aim was to compare these approaches for imputing three-level incomplete data specifically when the substantive analysis model includes interactions or quadratic effects involving incomplete covariates which need to be incorporated in the imputation model

Our focus was 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 motivating example aimed to estimate the effect of early depressive symptoms ($depression$) on the academic performance of the students ($NAPLAN$) over time adjusted for confounders, where there was incomplete data in the exposure of interest and at least one confounder i.e. multivariate missingness

Methods

Three analysis models that are common in longitudinal data settings were considered (with i denoting the i^{th}school, j denoting the j^{th} individual and k denoting the k^{th} wave) :

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

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave_{ijk} + \beta_3 \times depression_{ij(k-1)} \times wave_{ijk} + ** + b_{oi} + b_{oij} + \varepsilon_{ijk}$$

With $b_{oi} \sim N(0, \sigma^2_{boi}), b_{oij} \sim N(0, \sigma^2_{boij}) and\ \varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon ijk})$

2. A random intercept model with an interaction between the time-varying exposure and a time-fixed baseline variable (will not be discussed)

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

$$NAPLAN_{ijk} = \delta_0 + \delta_1 \times depression_{ij(k-1)} + \delta_2 \times wave_{ijk} + \delta_3 \times depression^2_{ij(k-1)} + ** + \alpha_{0i} + \alpha_{0ij} + e_{ijk}$$

With $\alpha_{oi} \sim N(0, \sigma^2_{\alpha oi}), \alpha_{oij} \sim N(0, \sigma^2_{\alpha oij}) and\ e_{ijk} \sim N(0, \sigma^2_{eijk})$

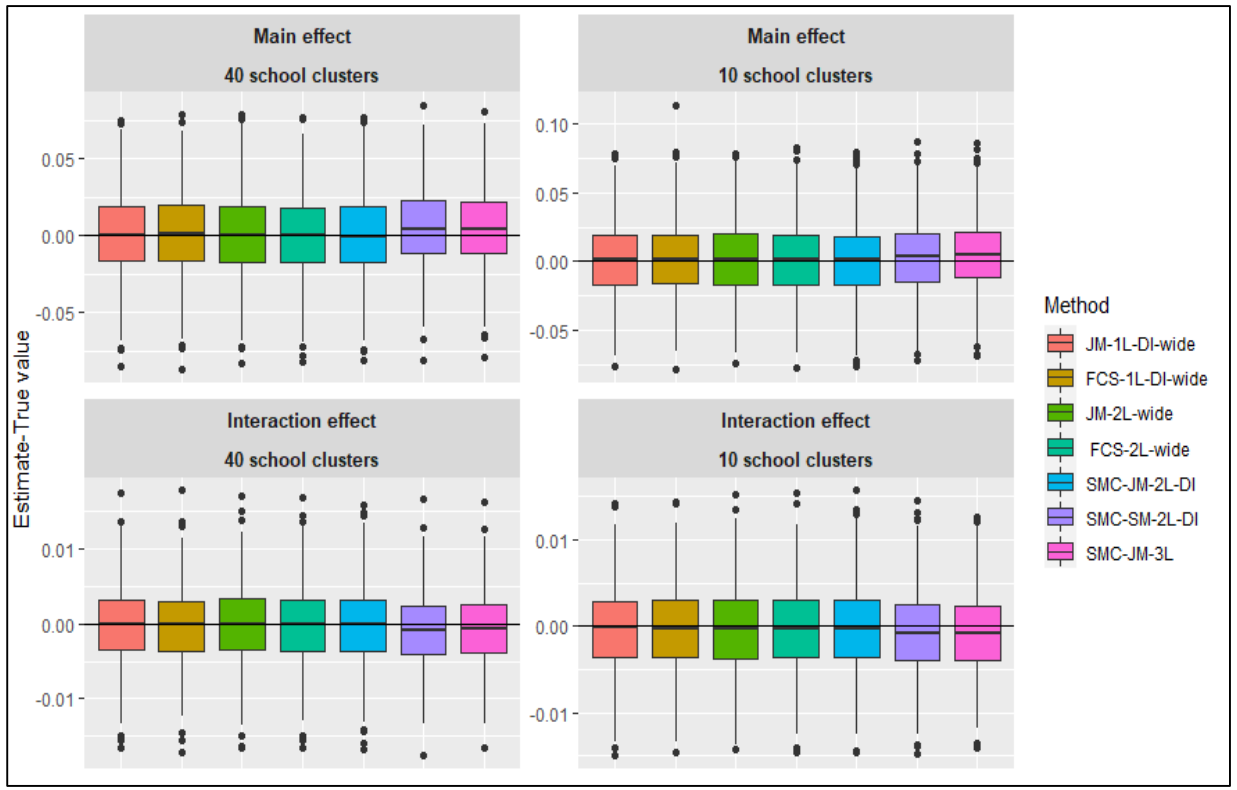
The data for the simulations were generated by closely mimicking the CATS data which was replicated 1000 times

Missing values were generated in the time-varying exposure of interest (15%, 20% and 30% of the depressive symptom scores at waves 2,4, and 6 respectively) according to a MAR mechanism and a time-fixed confounder (10 % of Socio-Economic Status) according to a MCAR mechanism.

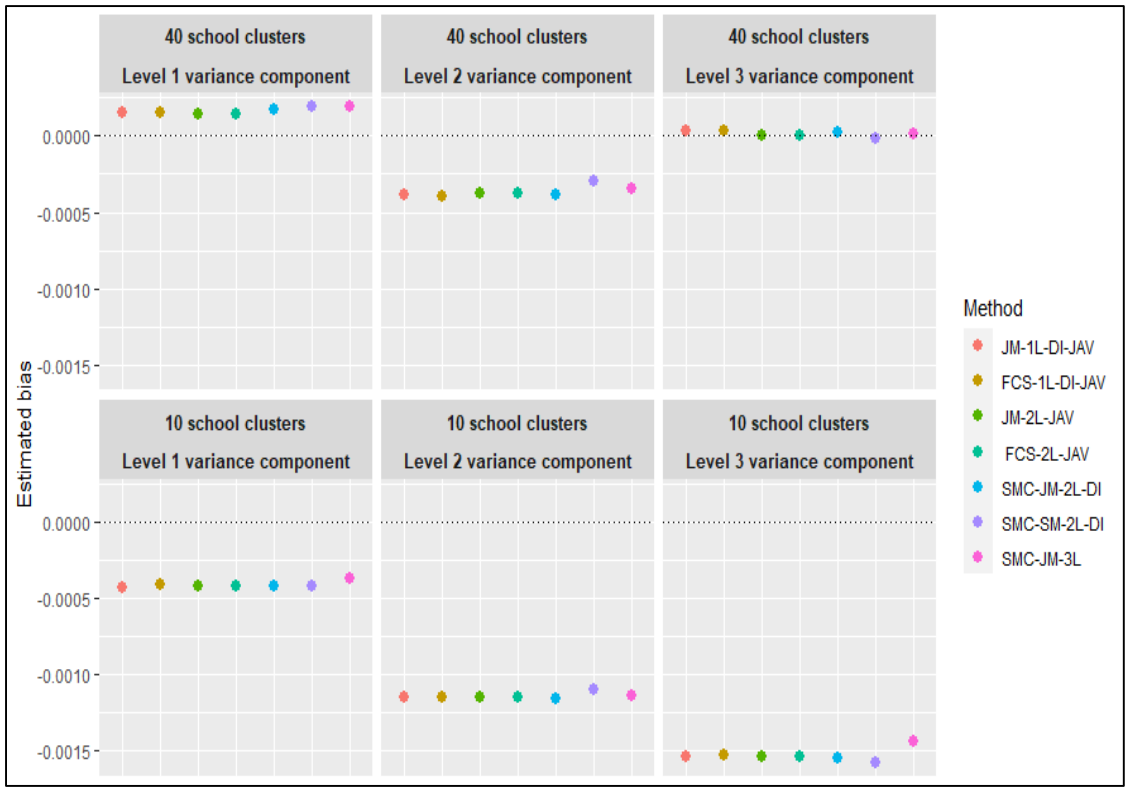
The behaviour of these approaches were also evaluated under two different numbers of higher level clusters: 40 school clusters and 10 school clusters

**the remaining covariates that were adjusted for in (1),(2) and (3) include Sex, SES, NAPLAN scores at wave 1 and Age at wave 1

Results-Analysis model 1



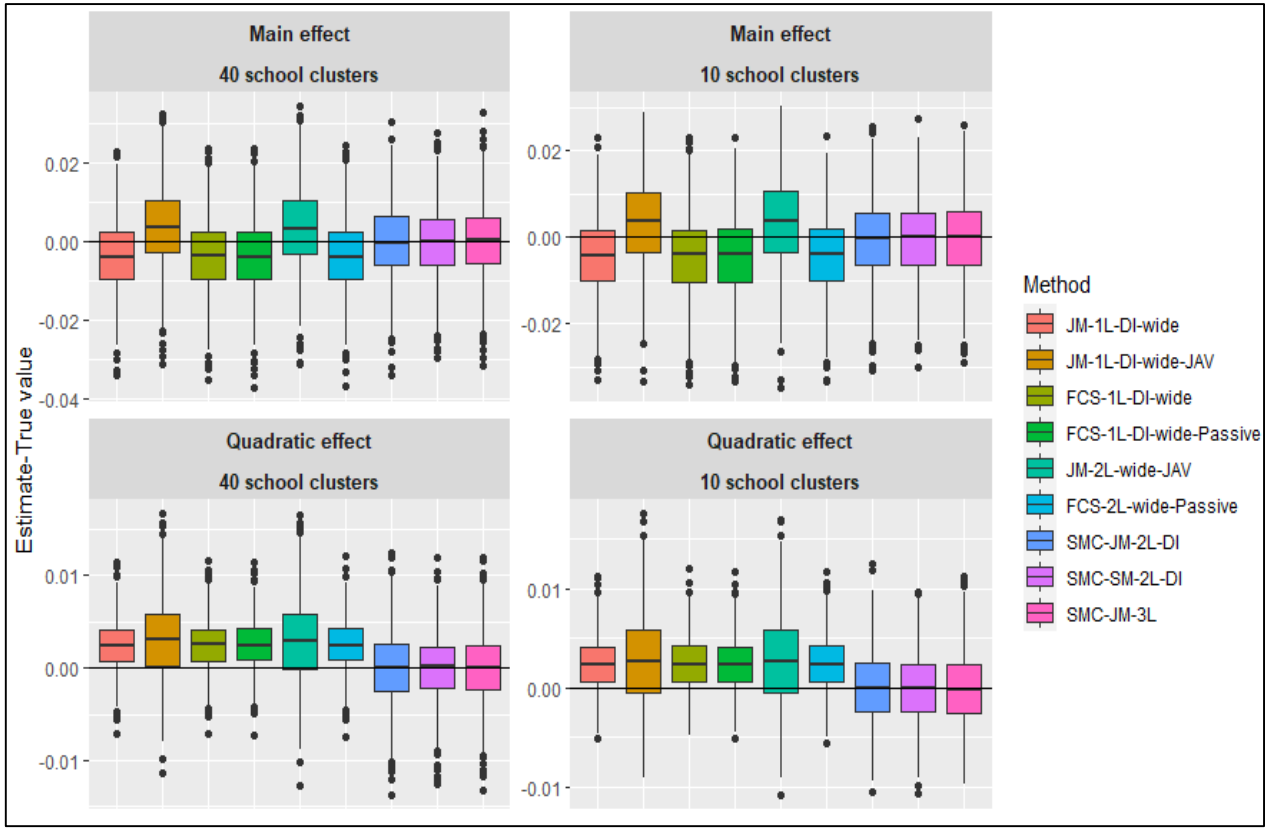
Distribution of the bias in the estimated regression coefficient for the main effect ($\beta_1, true\ value = -0.07$) and the interaction effect ($\beta_3, true\ value = 0.013$) across the 1000 simulated datasets for the 7 multiple imputation (MI) approaches



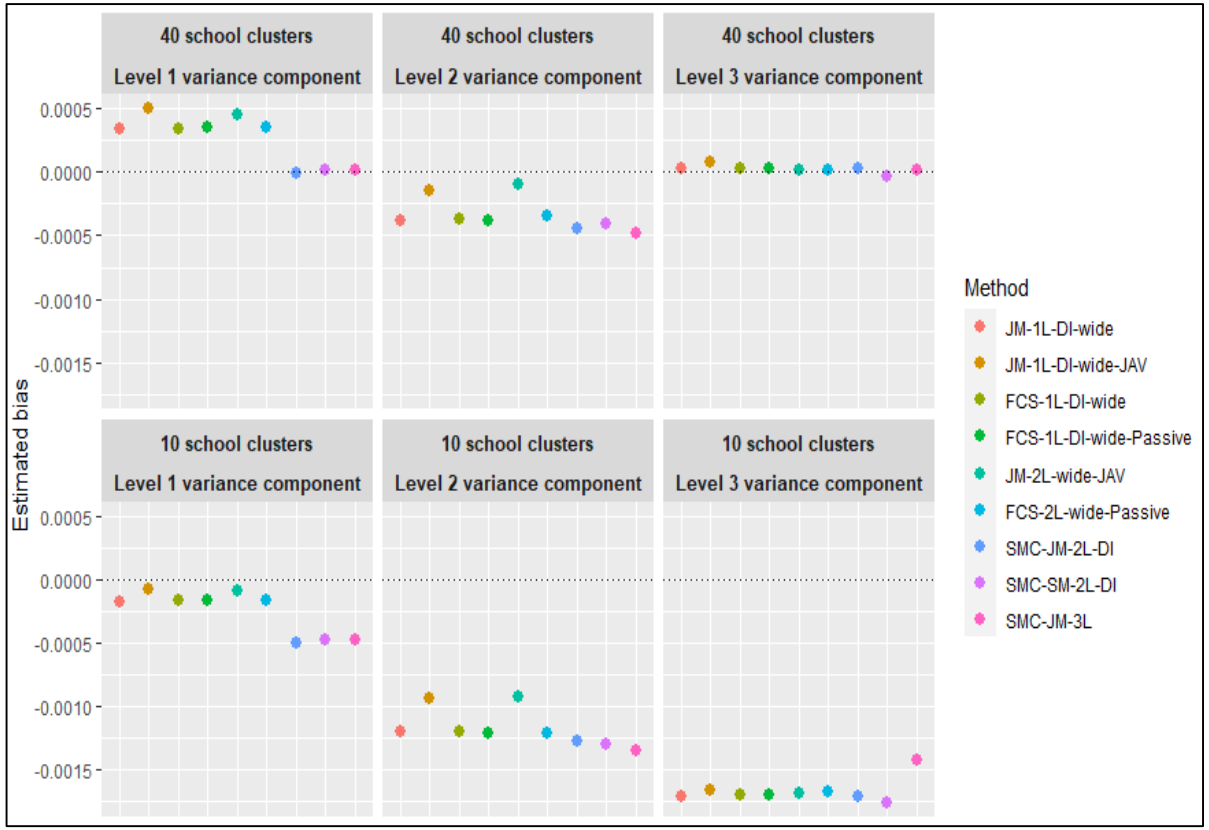
Estimated bias in the variance components at level 1, 2 and 3 in analysis model 1 across the 1000 simulated datasets for the 7 multiple imputation (MI) approaches

- All the MI approaches produced approximately unbiased estimates of the main effect and the interaction effect
- All approaches resulted in similar negligible bias (<10% relative bias) for the 3 variance components for both scenarios with slightly larger biases for the level 3 variance estimates when there were a smaller number of higher-level clusters

Results-Analysis model 3



Distribution of the bias in the estimated regression coefficient for the main effect ($\delta_1, true\ value = -0.024$) and the quadratic effect ($\delta_3, true\ value = -0.009$), across the 1000 simulated datasets for the 9 multiple imputation (MI) approaches



Estimated bias in the variance components at level 1, 2 and 3 across the 1000 simulated datasets for the 9 multiple imputation (MI) approaches

- All of the MI approaches except for ***SMC-JM-2L-DI, SMC-SM-2L-DI*** and ***SMC-JM-3L*** resulted in biased estimates for the main effect and the quadratic term, with substantial underestimation of the quadratic effect
- All approaches resulted in similar negligible bias (<10% relative bias) for the variance components across the different simulation scenarios, with slightly larger, bias for the level 3 and level 2 variance estimates when there were a smaller number of higher-level clusters

Conclusions

- With simple analysis models where there is only an interaction with time, all of the approaches (i.e. the three-level SMC MI approach and adaptations of single-and two-level approaches considered) are appropriate
- When the analysis model involves quadratic effects (or in general other interactions that do not involve time) the three-level SMC approach or the two-level SMC approaches with DI are more appropriate
- However under both the analysis models, approaches which uses the DI extension should be used with caution as it has been shown to produce biased parameter estimates in certain scenarios[7]
- Also the performance of these approaches are likely to vary when the analysis model involves random slopes[8]

(7) Drechsler J. Multiple imputation of multilevel missing data—rigor versus simplicity. J Educ Behav Stat. 2015;40(1):69–95.
(8) 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(2):444–66.