Author/s:
Mainzer, R;Apajee, J;Nguyen, CD;Carlin, JB;Lee, KJ

Title:
A comparison of multiple imputation strategies for handling missing data in multi-item scales: Guidance for longitudinal studies

Date:
2021-09-20

Citation:
Mainzer, R., Apajee, J., Nguyen, C. D., Carlin, J. B. & Lee, K. J. (2021). A comparison of multiple imputation strategies for handling missing data in multi-item scales: Guidance for longitudinal studies. Statistics in Medicine, 40 (21), pp.4660-4674. https://doi.org/10.1002/sim.9088.

Persistent Link:
https://hdl.handle.net/11343/298633

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH PAPER

A comparison of multiple imputation strategies for handling missing data in multi-item scales: guidance for longitudinal studies

Rheanna Mainzer*[1] | Jemishabye Apajee[1,2] | Cattram D. Nguyen[1,3] | John B. Carlin[1,4] | Katherine J. Lee[1,3]

[1]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia
[2]Quality Use of Medicines and Pharmacy Research Centre, Clinical and Health Sciences, University of South Australia, Adelaide, South Australia 5001
[3]Department of Paediatrics, The University of Melbourne, Parkville, Victoria 3052, Australia
[4]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3052, Australia

Correspondence
*Rheanna Mainzer, Murdoch Children's Research Institute, 50 Flemington Road, Parkville, Victoria 3052, Australia. Email: rheanna.mainzer@mcri.edu.au

Summary

Medical research often involves using multi-item scales to assess individual characteristics, disease severity and other health-related outcomes. It is common to observe missing data in the scale scores, due to missing data in one or more items that make up that score. Multiple imputation (MI) is a popular method for handling missing data. However, it is not clear how best to use MI in the context of scale scores, particularly when they are assessed at multiple waves of data collection resulting in large numbers of items. The aim of this paper is to provide practical advice on how to impute missing values in a repeatedly measured multi-item scale using MI when inference on the scale score is of interest. We evaluated the performance of five MI strategies for imputing missing data at either the item or scale level using simulated data and a case study based on four waves of the Longitudinal Study of Australian Children (LSAC). MI was implemented using both multivariate normal imputation (MVNI) and fully conditional specification (FCS), with two rules for calculating the scale score. A complete case analysis was also performed for comparison. Based on our results, we caution against the use of a MI strategy that does not include the scale score in the imputation model(s) when the scale score is required for analysis.

KEYWORDS:
multi-item scale, missing data, multiple imputation, longitudinal study

1 | INTRODUCTION

Medical research often involves using multi-item scales to assess individual characteristics, disease severity and other health-related outcomes. For example, the Paediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core scales is a 23-item scale that aims to measure health-related quality of life (HRQoL) in children.[1] The total summary score of this scale is derived from 23 individual scale items. Missing data arise in composite scale scores as a result of missing data in one or more items that

make up that score. Missing data in scale items can occur due to unit non-response, where individuals do not answer any of the items, or item non-response, where individuals answer some, but not all, of the items. Missing data in multi-item scales are often handled using a complete case analysis, where only individuals with data available on all variables of interest, which for a multi-item scale means all items, are included in the analysis, or single imputation methods, where either missing items or scale scores are filled in with a single imputed value and then the completed data are analysed.[2] However, both approaches have drawbacks. Removing incomplete cases from analysis leads to a loss of precision and, in many circumstances, biased parameter estimates,[3] and single imputation methods do not take into account the uncertainty surrounding imputed values.[4] To overcome these problems, one can use multiple imputation (MI) to handle missing data.[5] In the context of a multi-item scale, MI can be applied at either the item or scale level to obtain imputations for scale scores.

MI is a two-stage procedure. In the first stage, missing data are imputed multiple times from a specified imputation model based on observed data to obtain multiple completed datasets. In the second stage, each of these completed datasets is analysed separately using standard complete-data methods and these results are combined using Rubin's rules to produce an overall estimate of the parameter of interest and its corresponding standard error.[5,6] There are two approaches available in standard statistical software packages for imputing missing data in more than one variable: fully conditional specification (FCS) and multivariate normal imputation (MVNI). FCS, also known as "multiple imputation by chained equations" or "regression switching", specifies univariate imputation models, conditional on all other variables in the imputation model, for each variable with missing data.[7] Imputed values for each variable are generated one variable at a time using these models until all missing values are replaced (one cycle). This procedure is repeated several times to achieve convergence before obtaining one imputed dataset. The whole process is repeated multiple times to obtain multiple imputed datasets. Different imputation models can be specified for different variable types: for example, linear regression models for continuous variables and logistic regression models for binary variables. MVNI is a joint modelling approach that assumes all variables in the imputation model jointly follow a multivariate normal distribution. Imputations are obtained from this model using the Data Augmentation algorithm.[5] While MVNI is less flexible than FCS, it has been shown to perform well even when the multivariate normality assumption is violated.[8]

Standard implementations of MI are valid when data are missing at random (MAR). Under this assumption, missingness does not depend on the missing data once we have conditioned on the observed data,[9] although the precise definition of MAR can be difficult to interpret when there is missingness in more than one variable.[10,11] To make the MAR assumption more plausible, it is important to incorporate in the imputation model information about the observed data and the missing data mechanism. This is done by adding additional variables to the imputation model that are not in the analysis model, known as auxiliary variables. In particular, it is recommended to include in the imputation model (i) all variables in the analysis model, (ii) all variables correlated with the variable(s) with missing data, and (iii) all variables that are predictors of missingness .[5,7,12] Furthermore,

the imputation model should be congenial with the analysis model.[13,14] In its simplest form this means that all variables that appear in the analysis model should appear, in the same form, in the imputation model.

Several studies have investigated whether it is best to impute values at the item level or scale level when using MI to deal with missing data in multi-item scales. The general recommendation appears to be to impute at the item level (for gains in precision) prior to calculating the scale scores.[15,16,17] However, this is not always possible. For example, Rombach et al. encountered convergence problems when using FCS with ordinal logistic regression models to impute items in small samples.[18] In another study that considered two scales with missing items, simulation study results showed that imputing all items from both scales (item-level imputation) with MVNI had smaller bias and higher efficiency than either imputing scale totals or imputing each scale at the item level using the total scores of the other scale.[19] Nevertheless, this strategy gave rise to convergence problems for some imputation runs in a case study with real data, making its practical use questionable. Item-level imputation may fail due to numerical problems caused by perfect prediction or collinearity.[20] Such problems are commonly encountered in practice when imputing items in large-scale longitudinal studies because of the need to fit imputation models that contain a large number of highly correlated variables. Another paper noted that item-level imputation does not produce precision gains in longitudinal studies where entire questionnaire batteries are missing.[21] Furthermore, if the scale score is of interest in the analysis model but is left out of the imputation model, item-level imputation is not congenial with the analysis model, which could lead to bias.[7,13]

One way to overcome convergence problems caused by item-level imputation is to reduce the number of variables in the imputation model. In the context of several multi-item scales, it has been suggested to impute items using the rest of the items from the same scale and a summary (e.g., the mean or total score) of available items from other scales.[19,22] In simulation studies, this method has performed well compared with alternatives such as a complete case analysis or imputing total scores. Another way of reducing the number of variables in the imputation model is to use a variable reduction technique such as principal components analysis to produce summary measures of the scale items not used in the analysis model, which are then used as auxiliary variables in the imputation model. In a simplified simulation setting, Howard et al. showed that the principal components strategy performed similarly to an inclusive strategy that included all auxiliary variables in the imputation model, making it a promising alternative when the number of potential auxiliary variables is large or the inclusive strategy fails.[23] This approach has been embraced by applied researchers,[24,25,26] and implemented in statistical software.[27] However, its performance has been evaluated only in limited contexts.

Although some progress has been made to evaluate different MI strategies for imputing multi-item scales, most of this work has been done in the context of cross-sectional studies,[15,19,23,28] or randomised controlled trials with variables measured at baseline and follow-up.[18,22] Limited work has been done to evaluate MI strategies for imputing multi-item scales in the context of large-scale longitudinal studies where scales are measured across multiple waves of data collection. One such study by Nooraee et al. compared four MI strategies for handling missing data in 10 binary items measured at four waves.[17] However, the main aim of

that study was to evaluate a hybrid approach combining MI and maximum likelihood estimation, rather than to compare a range of possible MI strategies. Two other studies investigated two approaches for including item information as auxiliary variables when estimating a latent growth model: (1) including the items as auxiliary variables and (2) including a summary of the items (e.g., the mean of the observed items) as an auxiliary variable.[29,30] These studies found that incorporating item information as auxiliary variables led to estimates with greater precision compared with a complete case analysis or not including any auxiliary variables. However, these authors used full information maximum likelihood rather than MI and the number of approaches investigated was limited.

The aim of the current paper is to provide guidance on how to impute missing values using MI in a multi-item scale in the context of a large longitudinal study. We conducted a simulation study based on waves 1 to 4 of The Longitudinal Study of Australian Children (LSAC) to evaluate the performance of five different MI strategies for imputing missing data in repeatedly measured multi-item scales.[31] These strategies differ in the level at which the scale is imputed (item or scale level), as well as the auxiliary variables used in the imputation model. The MI strategies are (1) item-level imputation using items of scales measured at other waves as auxiliary variables, (2) scale-level imputation using scale scores from other waves as auxiliary variables, (3) item-level imputation using scale scores from other waves as auxiliary variables, (4) item-level imputation using the same items from other waves and scale scores at all waves as auxiliary variables in the imputation model and (5) item-level imputation using principal components derived from items of scales measured at other waves as auxiliary variables. A complete case analysis is performed for comparison. We also apply these strategies to an analysis of the LSAC data to illustrate their performance in practice. In Section 2 we describe the LSAC data, the design of the simulation study and the empirical example. Results are provided in Section 3, and a discussion of these results is provided in Section 4. Concluding remarks are given in Section 5.

2 | METHODS

2.1 | Illustrative example

LSAC is an ongoing large-scale longitudinal study, with the aim of investigating the effect of a child's environment on their well-being throughout life.[32] Two cohorts, each consisting of 5,000 children, were recruited into the study in 2003: the "B" (baby) cohort, who were aged 0–1 years at wave 1, and the "K" (kinder) cohort, who were aged 4–5 years at wave 1. The study is currently up to its ninth wave, with follow-up waves conducted every two years.

In this paper, we focus on a case study reported by Jansen et al.[33] One aim of this study was to investigate, using the first four waves of LSAC data in the K cohort, whether the cumulative burden of being overweight in early childhood was associated with reduced health-related quality of life (HRQoL) at age 10-11 years. The outcome variable of interest, HRQoL, was measured by the wave 4 Total Scale Score ("scale score") of the Generic Core Scales module of the Paediatric Quality of Life Inventory

(PedsQL).[1] These scales consist of 21 items for toddlers aged 2–4 years (the K cohort at wave 1) and 23 items for children and adolescents aged 5–18 years (the K cohort at waves 2–4). Each item has a 5-point Likert scale response, where a score of 1 corresponds to "never a problem" and a score of 5 corresponds to "always a problem". The scale score is calculated as the average of the observed scale items, after a reverse transformation of items in the following manner: 1=100, 2=75, 3=50, 4=25 and 5=0. Therefore, a higher scale score is indicative of a better HRQoL. The exposure variable of interest, cumulative overweight, was the number of waves (0 – 4) in which the child was categorised as either overweight or obese using age and gender-specific Body Mass Index (BMI) cutoffs.[34,35] BMI was calculated from height and weight measurements taken by an interviewer at each wave. The analysis was performed on a subset of 3898 children who had data on the outcome variable. FCS was used to impute missing values in the variables used to calculate cumulative burden and in the covariates used for confounder adjustment. This study found strong evidence of a positive association between the number of waves at which a child was overweight and poor future HRQoL.

To investigate the performance of different MI strategies for imputing missing values in a multi-item scale, we consider a simplified version of the analysis performed by Jansen et al.,[33] using data from the first four waves and all 4983 children in the LSAC K cohort. The exposure variable was BMI measured at wave 1 (which ranged from 10.7–27.9, with 1% missing). The outcome variable was HRQoL, measured by the PedsQL scale score at wave 4 (which ranged from 18.75–100). The PedsQL guidelines states that the scale score should not be calculated if more than 50% of the items in the scale are missing.[1] This resulted in 867 (17.4%) children with missing data in the outcome, i.e. the wave 4 HRQoL measure.

The effect of BMI on HRQoL was estimated using the linear regression model:

$$E(HRQoL) = \beta_0 + \beta_1 BMI + \beta_2 Female + \beta_3 Age + \beta_4 IndStat + \beta_5 NonEng + \beta_6 SEP, \quad (1)$$

where *HRQoL* is the scale score of the 23 PedsQL items at wave 4, *BMI* is the child's BMI measurement at wave 1, *Female* is an indicator for whether the child is female, *Age* is the age in months of the child at wave 1, *IndStat* is an indicator for whether the child is Aboriginal or Torres Strait Islander, *NonEng* is an indicator for whether the child has a non-English speaking background and *SEP* is the standardised socioeconomic position of the child's family at wave 1. The parameter of interest is β_1, the effect of *BMI* on *HRQoL*, controlling for the potential confounding effects of the covariates in the model. All variables in this analysis model had missing values except for *Age* and *Female*. If an individual had no missing data for any variable in the analysis model, they were a 'complete case'; otherwise they were an 'incomplete case'. Table 1 presents the number of missing values, and summary statistics for both the complete and incomplete cases, for each variable in the analysis model. Conducting a complete case analysis would mean that partial data from 927 participants (26%) would be discarded. Furthermore, differences between complete and incomplete cases in *SEP*, *IndStat* and *NonEng* suggested that a complete case analysis would result in biased estimates.

[Table 1 here]

2.2 | Simulation study

A simulation study based on the K cohort of LSAC was used to evaluate the performance of five different MI strategies for imputing missing data at either the item or scale level. These strategies also differ in how the information from other waves is included in the imputation model, i.e., which variables are used as auxiliary variables. We generated simulated data to replicate the 90 PedsQL items measured at waves 1 – 4 (items at waves 1 to 3 were used as, or to create, auxiliary variables), BMI measured at wave 1 and the Strengths and Difficulties Questionnaire measured at wave 4 (SDQ, range 0–33), which was used as an auxiliary variable. A total of 2000 complete datasets were generated so that the Monte Carlo standard error for an estimated coverage of 95% was less than 0.5%. Missingness was imposed under three different covariate dependent missingness (CDM) mechanisms, and three parameters of interest were estimated.

2.2.1 | Notation

Let $Y_{i,j}$ denote the j'th ordinal questionnaire item at wave i, based on the PedsQL items in the LSAC example, where $i = 1, 2, 3, 4$ and $j = 1, \ldots, 21$ if $i = 1$ and $j = 1, 2, \ldots, 23$ if $i = 2, 3, 4$. Let X denote a continuous predictor, based on BMI at wave 1, and Z denote a continuous auxiliary variable, based on SDQ at wave 4.

2.2.2 | Simulation of complete data

Each simulated dataset consisted of 1000 individuals, initially with complete data on the 92 variables $Y_{1,1}, \ldots, Y_{4,23}, X$ and Z. Data were generated as follows. We specified a mean vector with the first 90 elements equal to 0 and last two elements the estimates of the mean BMI at wave 1 and SDQ at wave 4 from the LSAC data. Similarly, we specified a standard deviation vector with the first 90 elements equal to 1 and last two elements the standard deviation estimates of BMI at wave 1 and SDQ at wave 4. The correlation matrix was estimated from the 90 PedsQL items, BMI at wave 1 and SDQ at wave 4. We drew 1000 observations from the multivariate normal distribution with this mean vector, standard deviation vector and correlation matrix. The first 90 variables were categorised to obtain observations for $Y_{1,1}, \ldots, Y_{4,23}$ as follows. The lowest 40% of ordered observations were allocated to category 1, the next 30%, 20%, 7% and 3% were allocated to categories 2, 3, 4 and 5, respectively. This categorisation was loosely based on the average proportion of items in each category for all observed PedsQL items in waves 1–4 for the K cohort in LSAC. We used the specified distribution so that the proportion of items in each category would sum to one and so that there was a slightly larger proportion of items in categories 4 and 5 than in LSAC (to avoid very small proportions in some categories). The categorisation was achieved using cut-points calculated from the inverse standard normal distribution, in a similar manner to Eekhout et al.[22]

2.2.3 | Imposing missing data

Missingness was imposed in the $Y_{i,j}$'s dependent on the auxiliary variable, Z, and the exposure of interest, X, under three different scenarios. For simplicity, missingness in items was assumed to occur independently of other items, and missingness in entire scales at each wave was assumed to occur independently of other waves. For each complete dataset, entire scales at each wave were set missing with a probability determined by the logistic regression model

$$\text{logit}\, P(\text{missing scales}) = \gamma_{\text{scale}} + \ln(\gamma_X)\, X^* + \ln(\gamma_Z)\, Z^*, \tag{2}$$

where X^* and Z^* denote the variables X and Z, respectively, rescaled to have unit variance, and the parameter vector $(\gamma_{\text{scale}}, \gamma_X, \gamma_Z) \in \mathbb{R}^3$ was used to specify the missingness scenario (described further below). Items for the remaining scales were then set missing with a probability determined by the logistic regression model

$$\text{logit}\, P(\text{missing items}) = \gamma_{\text{item}} + \ln(\gamma_X)\, X^* + \ln(\gamma_Z)\, Z^*, \tag{3}$$

where the parameter $\gamma_{\text{item}} \in \mathbb{R}$ was used to specify the missingness scenario. No missing data were generated in X or Z.

The three missingness scenarios considered were:

1. *CDM-LSAC:* The odds of missingness increased by 20% with every one standard deviation increase in X (Z), holding Z (X) fixed. Entire scales had a 15% chance of being missing at a given wave, and each item in the remaining scales had a 0.5% chance of being missing. These odds ratios were chosen based on the estimated odds of missingness in PedsQL items at waves 1 – 4, obtained by fitting a logistic regression model with the same form as model (2) to the LSAC data. The missingness proportions were chosen to reflect the missingness in LSAC. This scenario was achieved by setting $(\gamma_{\text{scale}}, \gamma_{\text{item}}, \gamma_X, \gamma_Y) = (-3.74, -7.31, 1.2, 1.2)$.

2. *CDM-inflated:* The odds of missingness increased by 20% with every one standard deviation increase in X (Z), holding Z (X) fixed. Entire scales had a 30% chance of being missing at a given wave, and each item in the remaining scales had a 1% chance of being missing. This scenario was included to explore the robustness of the MI analysis strategies under a larger proportion of missing data. It was achieved by setting $(\gamma_{\text{scale}}, \gamma_{\text{item}}, \gamma_X, \gamma_Y) = (-2.84, -6.58, 1.2, 1.2)$.

3. *CDM-extreme:* The odds of missingness increased by 100% with every one standard deviation increase in X (Z), holding Z (X) fixed. Entire scales had a 30% chance of being missing at a given wave, and each item in the remaining scales had a 1% chance of being missing. This scenario was included to explore the robustness of the MI analysis strategies under a larger proportion of missing data and a stronger CDM mechanism. It was achieved by setting $(\gamma_{\text{scale}}, \gamma_{\text{item}}, \gamma_X, \gamma_Y) = (-8.57, -12.31, 2, 2)$.

For each of the three missingness scenarios, the values of γ_{scale} and γ_{item} were chosen to be the values that obtained the desired proportion of missing data in a larger simulated dataset with $1,000,000$ individuals after setting the values for γ_X and γ_Z. This simulated dataset was also used to obtain parameters of interest (see Section 2.2.4) and tune the multiple imputation strategies (see Section 2.2.5).

2.2.4 | Parameters of interest

The analysis model was motivated by the illustrative example, but for simplicity there was assumed to be no confounding variables in the simulation study. In other words, the analysis model of interest in the simulation study was the linear regression model

$$Y_4 = \beta_0 + \beta_1 X, \tag{4}$$

where Y_4 denotes the scale score of the questionnaire items at wave 4. This score was calculated from $Y_{4,1}, Y_{4,2}, \ldots, Y_{4,23}$ in the same way as the PedsQL scale score (described earlier in Section 2.1). The parameters of interest were the coefficient β_1 in model (4), and the mean and median of Y_4. The "true" values of these parameters were obtained as the corresponding estimates from a large simulated dataset of 1,000,000 individuals using the Stata commands `regress` (for β_1), `mean` (for the mean of Y_4) and `qreq` (for the median of Y_4). For this dataset, $\beta_1 = -0.679$, the mean of Y_4 was 74.3 and the median of Y_4 was 76.1.

2.2.5 | Multiple imputation strategies evaluated

To use MI for analyses involving Y_4 we could either impute this variable directly (scale-level imputation) or impute the items $Y_{4,1}, Y_{4,2}, \ldots, Y_{4,23}$ and then calculate Y_4 from the observed and imputed items (item-level imputation). We compared the performance of the following five MI strategies:

1. *item-item:* Item-level imputation using items of scales measured at other waves as auxiliary variables in the imputation model(s).
2. *scale-scale:* Scale-level imputation using scale scores from other waves as auxiliary variables in the imputation model(s).
3. *item-scale:* Item-level imputation using scale scores from other waves as auxiliary variables in the imputation model(s).
4. *item-mix:* Item-level imputation using the same items from other waves and scale scores at all waves as auxiliary variables in the imputation models.
5. *item-PC:* Item-level imputation using principal components derived from items of scales measured at other waves as auxiliary variables in the imputation model(s).

For each method the imputation model(s) also included the predictor variable X and the auxiliary variable Z. The imputation model(s) for the *item-item* strategy consisted of 92 variables in total, of which 90 had missing data. For the *scale-scale* strategy there were 6 variables in total (4 with missing data). For the *item-scale* strategy there were 28 variables in total (26 with missing data). For the *item-mix* strategy, different imputation models were specified for each variable with missing data. The imputation models for items consisted of either 9 or 10 variables (depending on whether the PedsQL item in LSAC was measured at three or four waves) and the imputation models for scale scores consisted of the scale scores at other waves. For the *item-PC* strategy there were 32 variables in total (23 with missing data).

MI strategies *scale-scale*, *item-scale* and *item-mix* required scale scores to be calculated prior to imputation so they could be included in the imputation model. These scores were calculated using two rules: (i) the scale score is calculated only if all the items are observed, and (ii) following the PedsQL guidelines,[1] where the scale score is calculated as the average of the observed scale items if at least 50% of items are observed. An investigation of these rules, applied to the large simulated dataset after imposing missing data according to each missingness scenario, was carried out. Table 2 in the Supporting Information provides the percentage of individuals with all items observed, at least 50% of items observed and no items observed, for each wave and missingness scenario.

The *item-PC* strategy required the calculation of principal components from items measured at waves 1 to 3 prior to MI. The initial step of this strategy involved filling in missing values in these items so that principal component scores could be calculated for all individuals. Following Howard et al.,[23] we attempted to fill in these missing values using a single MVNI with all (wave 1 – 4) items, X and Z in the imputation model. However, as described ahead in Section 3.1, MVNI failed in 147 of the simulated datasets. So as not to impede the performance of the *item-PC* strategy, missing values in items at waves 1 – 3 were instead filled in with one FCS imputation (using all other items, X and Z in the imputation models) prior to obtaining principal components. We performed a PCA of the correlation matrix and retained the first seven principal components for use as auxiliary variables.[36] We chose the number of principal components using the large simulated dataset ($n = 1,000,000$) as follows. After setting values missing in this dataset according to each of the missingness scenarios in turn and then carrying out one initial MVNI imputation as described above (since MVNI did converge in the large dataset), scree graphs were produced.[37] Visual inspection of these graphs revealed that there was little difference in the magnitude of eigenvalues after the first seven principal components, with these seven principal components explaining roughly 42% of the variance in items at waves 1 - 3 (regardless of the missingness scenario). A similar percentage of variance was explained by the one principal component retained in the simulation study by Howard et al. (40% on average), and the seven principal components retained in the real data example (39.5%) presented by Howard et al.[23]

Each MI strategy was implemented using both FCS and MVNI, except for the *item-mix* strategy which could only be implemented using FCS. For the FCS implementation, univariate imputation models for variables with missing data included all other

variables specified by the strategy. For example, of the 90 variables with missing data in the *item-item* strategy, 23 were the wave 4 items used to calculate Y_4 and 67 were items measured at waves 1 - 3. The imputation models for each item included all 89 other items, X and Z. Both items and scale scores were imputed using predictive mean matching (PMM) with 10 candidate donors.[38] Initially, we considered imputing items using ordinal logistic regression models. However, this method encountered problems with perfect prediction and collinearity in the LSAC data, which made it unsuitable for use in practice and led us not to implement it in the simulation study. PMM was used rather than linear regression due to the non-normality of these item and scale scores. For the MVNI implementation, imputed items were not categorised or truncated following MVNI due to known issues with these approaches,[39,40,41] and also because the outcome of interest, Y_4, was treated as a continuous variable in the analysis (the items themselves were not of interest in the analysis model). For the item-level imputation strategies, Y_4 was obtained as the average of observed and imputed items following a reverse linear transformation, applied directly to the unrounded items, consistent with the PedsQL guidelines. The analysis was performed in Stata version 15,[42] using `mi impute chained` and `mi impute mvn` to implement FCS and MVNI, respectively.

Estimates, with associated standard errors and 95% confidence intervals, were obtained following MI using Rubin's rules with 40 imputed datasets. This was done using the Stata command `mi estimate` with either `regress` (for β_1), `mean` (for the mean of Y_4) or `qreq` (for the median of Y_4). These commands provided symmetric confidence intervals that were calculated using quantiles from the t distribution. The number of imputed datasets was chosen to be at least equal to the percentage of incomplete cases in the *CDM-extreme* scenario.[7] For consistency, the same number was used for the other two scenarios. A complete case analysis was carried out for comparison, where both rules (i) and (ii) were used to calculate scale scores. Estimates, standard errors and 95% confidence intervals for the complete case analysis were obtained using the Stata commands `regress`, `mean` and `qreg`.

2.2.6 | Performance criteria

Letting θ denote the "true" value of a parameter of interest (obtained from the large simulated dataset with $n = 1,000,000$), we let $\hat{\theta}_i$ and $se(\hat{\theta}_i)$ denote the estimate of θ and the standard error of the estimate of θ, respectively, from the i'th simulated dataset, where $i = 1, 2, \ldots, M$ and $M = 2000$ denotes the total number of simulated datasets. The performance of the 16 analysis strategies (five FCS approaches, four MVNI approaches and a complete case analysis, with two rules for computing scale scores where applicable) was compared for each parameter of interest (β_1, mean of Y_4, median of Y_4) for the *CDM-LSAC*, *CDM-inflated* and *CDM-extreme* scenarios using bias, calculated as $M^{-1}\sum_{i=1}^{M}\hat{\theta}_i - \theta$, empirical standard error (SE), calculated as the standard deviation of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$, the average model SE, calculated as the square root of $M^{-1}\sum_{i=1}^{M} se(\hat{\theta}_i)^2$, and the coverage probability of the 95% confidence interval, calculated as the proportion of times that θ was in the 95% confidence interval obtained from each simulated dataset.

A strategy with small bias and empirical SE compared with other strategies, and a coverage probability close to 0.95, is desirable. An average model SE close to the empirical SE indicates unbiased estimation of the model SE. Defining the standardised bias as (bias / empirical SE) × 100, we considered a strategy to perform poorly in terms of bias if the absolute value of the standardised bias exceeded 40%,[12] and in terms of coverage if the coverage obtained from the simulation was less than 94% (under-coverage) or more than 96% (over-coverage).

2.3 | Empirical example

We applied each of the 16 analysis strategies evaluated in the simulation study (described in the previous section) to the LSAC data to estimate the effect of BMI on HRQoL using model (1). Estimates and confidence intervals for β_1 were obtained using Rubin's rules and 40 imputed datasets.

The imputation model(s) for each analysis strategy included all seven variables in the analysis model (*HRQoL*, *BMI*, *Female*, *Age*, *IndStat*, *NonEng* and *SEP*), auxiliary variables required by the imputation strategy (either items of scales measured at waves 1 – 3, scale scores from waves 1 – 3, the same item from waves 1 – 3 and scale scores from waves 1 – 3, or principal components that were derived from items of scales measured at waves 1 – 3), and repeated measures on the following five additional auxiliary variables: *Global Health Measure*, a rating of the child's current health (*GHM*, range 1 – 5, measured at waves 1 – 4), *Special health care needs*, a variable that indicates whether the child had a health condition that required special care (*SHCN*, 0/1, measured at waves 1 – 4), *Strengths and Difficulties questionnaire*, a scale score providing information on the child's behaviour (*SDQ*, range 0 – 40, measured at waves 1 – 4), *Matrix Reasoning test*, a measure of the non-verbal intelligence of the child (*MR*, range 1 – 19, measured at waves 2 – 4) and *Peabody Picture Vocabulary Test*, a scale score providing information on the child's vocabulary (*PPVT*, range 28–106, measured at waves 1 –3). These variables were chosen because they were either correlated with *HRQoL* or *BMI*, or they were associated with missingness in *HRQoL* or *BMI*. Further details of these associations are provided in Table 1 of the online Supporting Information.

All analysis strategies were implemented in the same way as the simulation study, with the following additional details specific to this example. Rules (i) and (ii) for calculating the scale score were applied to the PedsQL items only. To avoid extra complexity, these rules were not applied to the additional auxiliary variables that were scale scores. Instead, if these additional auxiliary variables had missing values they were imputed at the scale level using the derived scale score provided as part of the LSAC data. When using FCS, linear regression models (`regress`) were used to impute *BMI*, *SEP*, *MR*, *SDQ*, *PPVT* and *GHM*, predictive mean matching with 10 candidate donors (`pmm`) was used to impute PedsQL items and scale scores, and logistic regression models (`logit`) were used to impute *IndStat*, *NonEng* and *SHCN*. When using MVNI, binary variables in the analysis model were dichotomised following the imputation procedure using adaptive rounding, which rounds imputed values to 0 or 1 using a threshold based on a normal approximation to the binomial distribution.[43]

3 | RESULTS

3.1 | Simulation study

Figures 1 and 2 present the results for the *CDM-LSAC* and *CDM-extreme* scenarios, respectively. The top panel of these figures shows the bias for each of the parameters of interest (left to right: β_1, mean of Y_4, median of Y_4), the middle panel shows the empirical SE and the bottom panel shows the coverage. Dashed horizontal reference lines are added to the plots of bias at 0 and 10% of the true parameter value, and to the coverage at 94% and 96% to provide guidance for the reader. Results from the complete case analysis are included in these figures as a benchmark for what can be expected when no MI is used. Results from the *CDM-inflated* scenario did not differ greatly from the *CDM-LSAC* and *CDM-extreme* scenarios, and are therefore provided in Figure 1 of the online Supporting Information. However, one notable distinction between the *CDM-inflated* scenario and the other two scenarios was that, when the *item-item* strategy was implemented with MVNI in the *CDM-inflated* scenario, the MI algorithm failed in 147 out of 2000 datasets (7%). Thus the results presented for this strategy in the Supporting Information are for the remaining 1853 simulated datasets. Because these results may present an overly optimistic picture of the performance of this strategy, they should be interpreted with caution. Tables 3 to 5 in the online Supporting Information present full results for each scenario, with summaries of the Monte Carlo SEs in the bias and empirical SE to quantify simulation uncertainty.[44] In the following subsections we highlight the key findings for the *CDM-LSAC* and *CDM-extreme* scenarios.

3.1.1 | CDM-LSAC

All MI strategies considered led to estimates of β_1 with bias less than 10% of the true parameter value and adequate coverage probability. The largest biases for β_1 were obtained when the *item-item*, *item-scale* and *item-PC* strategies (which impute the scale items) were implemented using FCS. These values were comparable with the bias obtained from the complete case analysis. A similar pattern was observed when the parameter of interest was the mean of Y_4. The largest empirical SEs for both β_1 and the mean of Y_4 were obtained from the *scale-scale* strategy, when scale scores were calculated prior to MI only if all items were observed (rule i). However, when scale scores were calculated if at least 50% of items were observed (rule ii), the empirical SE from this strategy was comparable to the empirical SEs from the other MI strategies. The two rules for calculating scale scores did not appear to have a substantial impact on either the *item-scale* strategy or the *item-mix* strategy. The 95% confidence intervals for the mean of Y_4 obtained from the *item-item*, *item-scale* and *item-PC* strategies using FCS had lower than nominal coverage, with the *item-item* strategy producing confidence intervals that contained the true mean in just 88% of simulated datasets.

In terms of the median of Y_4, each MI strategy resulted in estimates with bias less than 10% of the true parameter value. However, strategies implemented with MVNI underestimated the median of Y_4, on average, while the *item-item*, *item-scale* and *item-PC* strategies implemented with FCS overestimated the median of Y_4, on average. The largest empirical SE was obtained

from the *scale-scale* strategy using FCS and rule i. When rule ii was used instead, this empirical SE reduced slightly but was still larger than the empirical SEs obtained from the other MI strategies. Only the *scale-scale* strategy using MVNI produced confidence intervals with adequate coverage probability. However, this strategy had the largest negative bias. When implemented with FCS, the *item-item* strategy produced confidence intervals with lower than nominal coverage (93%), while all other MI strategies produced confidence intervals that contained the true median in more than 96% of the simulated datasets.

[Figure 1 here]

3.1.2 | CDM-extreme

This scenario had both a larger proportion of missing data and a stronger CDM mechanism than the *CDM-LSAC* scenario. All analysis strategies (including the complete case analysis) led to larger biases than the *CDM-LSAC* scenario. However, this increase in bias was relatively larger for the complete case analysis than the MI strategies. Overall, the pattern of results was similar to the *CDM-LSAC* scenario. That is, the largest bias for β_1 and the mean of Y_4 were obtained when the *item-item*, *item-scale* and *item-PC* strategies were implemented using FCS. Strategies that used MVNI tended to underestimate the bias for the median of Y_4, while those that used FCS tended to overestimate it. Again, the largest empirical SE was obtained from the *scale-scale* strategy using rule i. When rule ii was used instead, empirical SEs obtained from this strategy were comparable with the empirical SEs obtained from the other MI strategies. When implemented with FCS, the *item-item* strategy led to 95% confidence intervals with the lowest coverage probability for all three parameters (78% for β_1, 30% for the mean of Y_4 and 73% for the median of Y_4). When implemented with FCS, the *item-mix* and *item-PC* strategies also led to 95% confidence intervals for the mean of Y_4 with very low coverage probability (62% for *item-mix* using rule i, 64% for *item-mix* using rule ii, and 55% for *item-PC*).

[Figure 2 here]

3.2 | Empirical example

Figure 3 presents estimates, with 95% confidence intervals, for the parameter β_1 in model (1), i.e., the effect of BMI on HRQoL for the LSAC illustrative example, obtained from each of the MI strategies. Results from the analysis of complete cases are presented for comparison. Estimates from the complete case analysis were closer to 0 than estimates obtained using MI. The *scale-scale* strategy led to the widest confidence intervals for both FCS and MVNI when scale scores were calculated prior to MI if all the items were observed (rule i). However, these confidence intervals became narrower when scale scores were calculated following the PedsQL guidelines (rule ii), consistent with the results of the simulation study. Although there were slight differences between the estimates from the different MI strategies, they all led to very similar conclusions. That is, there

is strong evidence that an increase in BMI in Australian children aged 4–5 (age of children at wave 1 of LSAC) is associated with a decrease in HRQoL at age 10–11 (age of children at wave 4 of LSAC) of around 0.7 standardised units per unit of BMI.

[Figure 3 here]

4 | DISCUSSION

In this paper, we compared the performance of five MI strategies, implemented using both MVNI and FCS and with two rules for computing scale scores, for imputing missing values in a multi-item scale when inference on the scale score is of interest. When the parameter of interest was either the mean or regression coefficient, estimates with larger bias and 95% confidence intervals with lower than nominal coverage were obtained when item-level imputation strategies that did not include the scale score in the imputation model were implemented with FCS using PMM (strategies *item-item*, *item-scale* and *item-PC*). When there was a larger proportion of missing data and stronger CDM mechanism, these problems with bias and coverage (for the item-level imputation strategies using FCS) became worse. In the *CDM-inflated* missingness scenario, the *item-item* strategy (scale scores imputed at the item-level using items from all waves as auxiliary variables) encountered convergence problems when implemented with MVNI. All other MI strategies led to reasonable bias and coverage for the mean and regression coefficient parameters. When the parameter of interest was the median scale score, similar patterns of bias were seen from the *item-item*, *item-scale* and *item-PC* strategies, while all strategies implemented with MVNI underestimated the median. Overall, MI performed better than the complete case analysis in this simulation study. For the complete case analysis, the bias for each of the parameters increased as the proportion of missingess and strength of missingess mechanism increased. This led to values of the coverage that were far below nominal in the *CDM-extreme* scenario. The bias for β_1 may be unexpected given the missingess was not related to the outcome. However, this can be explained by the inclusion of the auxiliary variable, Z, which was not in the analysis model but was related to the missingness and the outcome, and therefore induced a correlation between missingness and the outcome. The findings of the LSAC empirical example generally reflected those of the simulation study, in particular the *CDM-LSAC* scenario. That is, most MI strategies led to similar estimates; estimates from the complete case analysis were noticeably different to the estimates from the MI strategies; and the *scale-scale* strategy using rule (i) led to estimates with the largest standard errors. The only difference was that there was no noticeable distinction between FCS and MVNI in the empirical example.

Previous simulation studies performed with smaller sample sizes have recommended item-level imputation over scale-level imputation for incomplete scale scores in order to improve precision.[15,22] We also observed higher precision resulting from item-level imputation. However, when implemented with FCS and PMM, we found larger bias in estimates produced from item-level imputation strategies that did not include the scale score in the imputation models. To obtain valid results from MI, the

imputation model should be congenial with the analysis model. In our simulation study, this would require Y_4 to be included in the imputation model. Since Y_4 is a linear combination of the items at wave 4, we cannot fit the model including all of these variables simultaneously (parameter estimates can not be obtained due to collinearity). Therefore, we would expect to see poorer performance from the *item-item*, *item-scale* and *item-PC* strategies. This was observed when these MI strategies were implemented with FCS and PMM, but not MVNI. In the current study, we also found that applying MI at the scale level (*scale-scale*) led to estimates with large empirical SE relative to the other MI strategies when scale scores were calculated prior to MI only if all items were observed. However, when scale scores were calculated if more than 50% of items were observed, applying MI at the scale level led to empirical SEs that were similar to those obtained from applying MI at the item level. The fact that scale-level imputation appeared to perform well in the current setting is consistent with other studies that show item and scale-level approaches are comparable in large samples.[18,28] In practice there are many possible approaches for incorporating item information in MI, for example one could use the average of all observed items at each wave as auxiliary variables. Consistent with the results of Eekhout et al.,[30] we found that incorporating more item information (using rule ii instead of rule i) improved the precision of estimates when imputation was conducted at the scale level. However, when imputation was conducted at the item level, we observed little difference between the two rules. Our decision to investigate these particular two rules was based on the guidelines for the PedsQL questionnaire. Nevertheless, results would not have changed much if we calculated scale scores using all available information since not many participants had between 0 and 50% of items observed.

Using principal components analysis to reduce the set of auxiliary variables to a reasonable number, whilst still retaining most of the variation among the candidate auxiliary variables is an appealing strategy. We found that item-level imputation using principal components derived from items of scales measured at other waves as auxiliary variables in the imputation model (*item-PC*) was an acceptable strategy for dealing with multi-item scales measured at multiple waves when implemented with MVNI. (As described earlier, when FCS with PMM was used instead of MVNI, this strategy resulted in estimates with larger bias and worse coverage than many of the other MI strategies we considered.) In order to implement this strategy, we had to choose the number of principal components to use and how to fill in missing items prior to obtaining principal components. Additional research is needed to provide guidance around these choices. In our simulation study, using a single MVNI with all items in the imputation model to obtain an initial single imputation as suggested by Howard et al. failed in a number of simulated datasets in the *CDM-inflated* scenario.[23] Others have proposed alternative uses of principal components in the handling of missing data,[45] although this is currently only proposed as a single imputation technique. Furthermore, principal components analysis is just one of many dimension reduction or variable selection techniques that could be applied to specify an imputation model.[46] Future research comparing such approaches is needed.

In this paper, we imputed items and scale scores using FCS with PMM because this approach adequately captured the non-normal distribution of these variables. However, we could have instead used ordinal logistic regression (to impute items) or

linear regression (to impute items and scale scores). We repeated the simulation study using linear regression models to impute incomplete items and scale scores, but this led to an unrealistic symmetric distribution of both the imputed scale scores and the scale scores derived from imputed items, and, as expected, performed very similarly to MVNI. (MVNI is equivalent to FCS when linear regression models are specified to impute each variable.) We also tried using ordinal logistic regression to impute items in the LSAC empirical example, but we encountered problems with perfect prediction and collinearity. In order for the algorithm to proceed, problematic variables were automatically dropped from the imputation model, making it different from the specified model and unsuitable to evaluate in our simulation study. Others have documented similar poor convergence with FCS when using ordered logistic regression models to impute incomplete items.[18] Our study also illustrates the difficulty of estimating a percentile using MI when the distribution of the incomplete variable is non-normal.[14] If the parameter of primary interest is the median scale score, it may be best to consider transformations of items or scale scores prior to MI.

The current study addresses an important gap in the MI literature regarding how to best impute missing values in a multi-item scale assessed at multiple waves of data collection. The simulation study was designed to be as realistic as possible, using data from the first four waves of LSAC. In addition to the five strategies considered, we also investigated the choice of MI method (FCS or MVNI) and two rules for calculating scale scores across three different CDM scenarios, making it a comprehensive comparison. There are, however, several limitations of this study that should be noted. Firstly, missingness was imposed in each item independently of other items, and in entire scales at each wave independently of other waves. Also, missingness in items depended on the same covariates as missingness in entire scales. In reality, these assumptions are unlikely to hold. For example, missingness in items is likely to depend on whether or not other items were answered. Nevertheless, this study provides a natural starting point for evaluating the available MI strategies. Secondly, the simulation study was based on a particular longitudinal study, so findings may not generalise to other studies. However, this is inevitable for any simulation study that attempts to mimic a realistic analysis setting, and the target study here has many features that occur commonly in practice. The simulation study was also limited in terms of the data generating process, for example, using the same cut-offs to categorise items, and the scenarios considered, for example, we did not investigate how the MI estimates from each strategy were affected by sample size or correlation between items, and the analysis model in the simulation study had missing data only in the outcome variable. In practice (as in the illustrative example) missing data are likely to arise in several variables. However, disentangling effects in more complex settings is difficult and there is also no reason to expect different results. Finally, we considered the missingness mechanism to be dependent on the covariates, whereas in practice the missingness mechanism may be more complex. However, this opens up the much broader topic of sensitivity analysis, which is beyond the scope of the current investigation.

5 | CONCLUSION

We evaluated a range of possible MI strategies for imputing missing data at either the item or scale level of a multi-item scale which is measured at multiple waves of follow-up, with the aim of providing practical advice for researchers faced with this problem, a topic for which little guidance is currently available. In this study we found that: (1) when strategies were implemented with FCS and PMM, not including the scale score of interest in the imputation models resulted in biased estimates, (2) MVNI models that included items at all waves encountered convergence problems, and (3) scale-level imputation performed well when the scale score was calculated as the average of the observed scale items if at least 50% of items were observed, regardless of whether this strategy was implemented with MVNI or FCS. Based on our results, we caution against the use of a MI strategy that does not include the scale score in the imputation model(s) when the scale score is required for analysis.

ACKNOWLEDGMENTS

This paper uses unit record data from Growing Up in Australia: the Longitudinal Study of Australian Children (LSAC). LSAC is conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the authors and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: 10.26193/F2YRL5.

Availability of data

The data that support the findings of this study are available from the DSS. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/F2YRL5 with the permission of the DSS.

Availability of code

The code used to produce the simulated datasets and carry out the analysis for the simulation study are available on GitHub: https://github.com/rheanna-mainzer/MI-multi-item-scales.

Author contributions

All authors provided input into the study design. JA and RM wrote the Stata programs for the LSAC analysis and simulation study. RM wrote the manuscript with input from KL. All authors provided feedback on the manuscript.

Financial disclosure

This work was supported by an Australian National Health and Medical Research Council (NHMRC) Career Development Fellowship (CDF) Level 2 Grant (grant 1127984) and a NHMRC Project Grant (grant 1166023). Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Varni JW, Burwinkle TM, Seid M, Skarr D. The PedsQL™* 4.0 as a pediatric population health measure: feasibility, reliability, and validity. *Ambul Pediatr* 2003; 3(6): 329–341.
2. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology* 2012; 23(5): 729–732.
3. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010; 29(28): 2920–2931.
4. Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: Wiley. 2 ed. 2002.
5. Schafer JL. *Analysis of incomplete multivariate data*. New York: Chapman & Hall . 1997.
6. Rubin DB. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons . 2004.
7. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30(4): 377–399.
8. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol* 2010; 171(5): 624–632.
9. Rubin D. Inference and missing data. *Biometrika* 1976; 63(3): 581–592.
10. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *Am J Epidemiol* 2018; 187(12): 2705–2715.
11. Seaman S, Galati J, Jackson D, Carlin J. What is meant by "Missing at Random"?. *Stat Sci* 2013: 257–268.

12. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; 6(4): 330.

13. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 1994: 538–558.

14. Van Buuren S. *Flexible imputation of missing data*. CRC press . 2018.

15. Eekhout I, de Vet HCW, Twisk JWR, Brand JPL, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol* 2014; 67(3): 335–342.

16. Gottschall AC, West SG, Enders CK. A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behav Res* 2012; 47(1): 1–25.

17. Nooraee N, Molenberghs G, Ormel J, Van den Heuvel ER. Strategies for handling missing data in longitudinal studies with questionnaires. *J Stat Comput Simul* 2018; 88(17): 3415–3436.

18. Rombach I, Gray AM, Jenkinson C, Murray DW, Rivero-Arias O. Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Med Res Methodol* 2018; 18(1): 87.

19. Plumpton CO, Morris T, Hughes DA, White IR. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Res Notes* 2016; 9(1): 45.

20. Nguyen CD, Carlin JB, Lee KJ. Practical strategies for handling numerical difficulties with multiple imputation algorithms. *Emerg Themes Epidemiol* 2021.

21. Vera JD, Enders CK. Is item imputation always better? An investigation of wave-missing data in growth models. *Struct Equ Modeling* 2021: 1–12.

22. Eekhout I, de Vet HCW, de Boer MR, Twisk JWR, Heymans MW. Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales. *Stat Methods Med Res* 2018; 27(4): 1128–1140.

23. Howard WJ, Rhemtulla M, Little TD. Using principal components as auxiliary variables in missing data estimation. *Multivariate Behav Res* 2015; 50(3): 285–299.

24. Erentaitė R, Vosylis R, Gabrialavičiūtė I, Raižienė S. How does school experience relate to adolescent identity formation over time? Cross-lagged associations between school engagement, school burnout and identity processing styles. *J Youth Adolesc* 2018; 47(4): 760–774.

25. Jensen AC, McHale SM, Pond AM. Parents' social comparisons of siblings and youth problem behavior: A moderated mediation model. *J Youth Adolesc* 2018; 47(10): 2088–2099.

26. Latzman NE, Vivolo-Kantor AM, Niolon PH, Ghazarian SR. Predicting adolescent dating violence perpetration: Role of exposure to intimate partner violence and parenting practices. *Am J Prev Med* 2015; 49(3): 476–482.

27. Lang KM, Little TD, Chesnut S, Gupta V, Jung B, Panko P. PcAux: Automatically extract auxiliary features for simple, principled missing data analysis [R Package]. 2017.

28. Simons CL, Rivero-Arias O, Yu LM, Simon J. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index?. *Qual Life Res* 2015; 24(4): 805–815.

29. Eekhout I, Enders CK, Twisk JW, De Boer MR, De Vet HC, Heymans MW. Including auxiliary item information in longitudinal data analyses improved handling missing questionnaire outcome data. *J Clin Epidemiol* 2015; 68(6): 637–645.

30. Eekhout I, Enders CK, Twisk JW, De Boer MR, De Vet HC, Heymans MW. Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Struct Equ Modeling* 2015; 22(4): 588–602.

31. Department of Social Services, Australian Institute of Family Studies and Australian Bureau of Statistics . Growing Up in Australia: Longitudinal Study of Australian Children (LSAC) Release 7.2 (Waves 1-7). 2020

32. Sanson A, Nicholson J, Ungerer J, et al. Introducing the longitudinal study of Australian children-LSAC discussion paper no. 1. *Melbourne, Australia: Australian Institute of Family Studies* 2002.

33. Jansen PW, Mensah FK, Clifford S, Nicholson JM, Wake M. Bidirectional associations between overweight and health-related quality of life from 4–11 years: Longitudinal Study of Australian Children. *Int J Obes* 2013; 37(10): 1307–1313.

34. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* 2000; 320(7244): 1240.

35. Cole TJ, Flegal KM, Nicholls D, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ* 2007; 335(7612): 194.

36. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 2016; 374(2065): 20150202.

37. Cattell RB. The scree test for the number of factors. *Multivariate Behav Res* 1966; 1(2): 245–276.

38. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014; 14(1): 75.

39. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: a simulation study. *BMC Med Res Methodol* 2014; 14(1): 57.

40. Lee KJ, Galati JC, Simpson JA, Carlin JB. Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Stat Med* 2012; 31(30): 4164–4174.

41. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *Am Stat* 2003; 57(4): 229–232.

42. StataCorp . Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017.

43. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med* 2007; 26(6): 1368–1382.

44. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; 38(11): 2074–2102.

45. Audigier V, Husson F, Josse J. A principal component method to impute missing values for mixed data. *Adv Data Anal Classif* 2016; 10(1): 5–26.

46. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. 112. Springer . 2013.

Variable	Missing values *n* (%)	Complete cases Mean (SD) / *n* (%)	Incomplete cases Mean (SD) / *n* (%)
Continuous			
HRQoL	867 (17.4)	77.6 (14.4)	72.3 (17.7)
BMI	49 (0.98)	16.32 (1.64)	16.42 (2.09)
Age	0 (0)	56.9 (2.6)	57.0 (2.7)
SEP	18 (0.36)	0.1 (0.98)	-0.44 (0.96)
Binary			
Female	0 (0)	1988 (49.01)	459 (49.51)
IndStat	2 (0.04)	108 (2.66)	79 (8.54)
NonEng	37 (0.74)	524 (12.92)	254 (28.54)

TABLE 1 Summaries of the amount of missing data, and the observed data for the complete cases and incomplete cases, for each variable in the analysis model. The complete cases consist of 4056 individuals with no missing data for any of the variables in the analysis model and the incomplete cases consist of the remaining 927 individuals (summaries based on available data). Mean (standard deviation [SD]) are presented for continuous variables or number with characteristic (%) for binary variables.

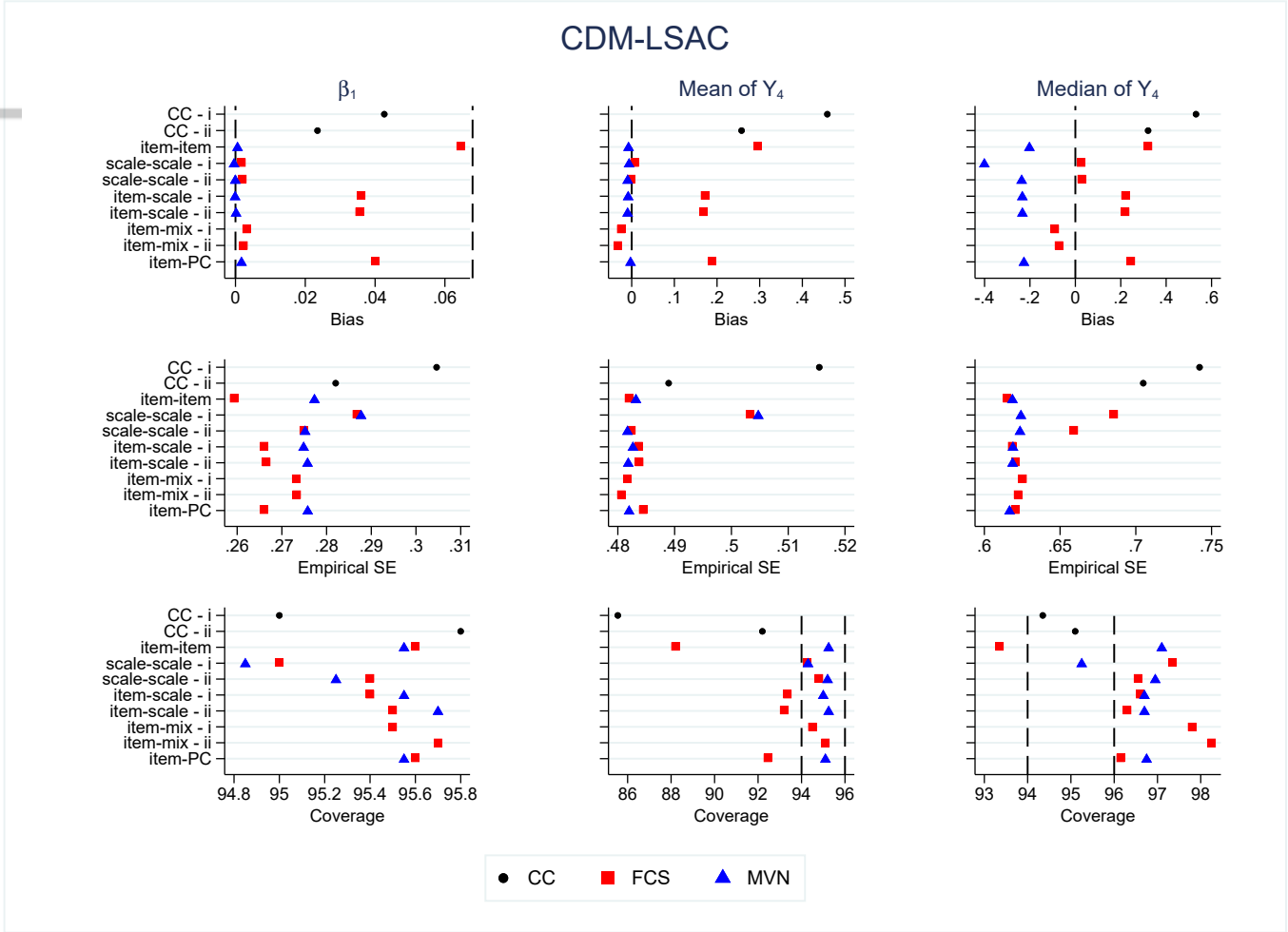

FIGURE 1 Bias, empirical standard error (SE) and coverage probability of the 95% confidence intervals for β_1, mean of Y_4 and median of Y_4, for the *CDM-LSAC* scenario. Analysis was performed using either a complete cases analysis (CC), multiple imputation (MI) by fully conditional specification (FCS) or multivariate normal imputation (MVN). Five MI strategies were considered: *item-item*, item-level imputation using other items as auxiliary variables; *scale-scale*, scale-level imputation using other scale scores as auxiliary variables; *item-scale*, item-level imputation using scale scores as auxiliary variables; *item-mix*, item-level imputation using a mix of items and scales as auxiliary variables; *item-PC*, item-level imputation using principal component scores as auxiliary variables. Where applicable, scale scores calculated according to one of two rules (i and ii). Dashed horizontal reference lines are added to the bias figures at 0 and 10% of the true parameter value, and the coverage figures at 94% and 96%.

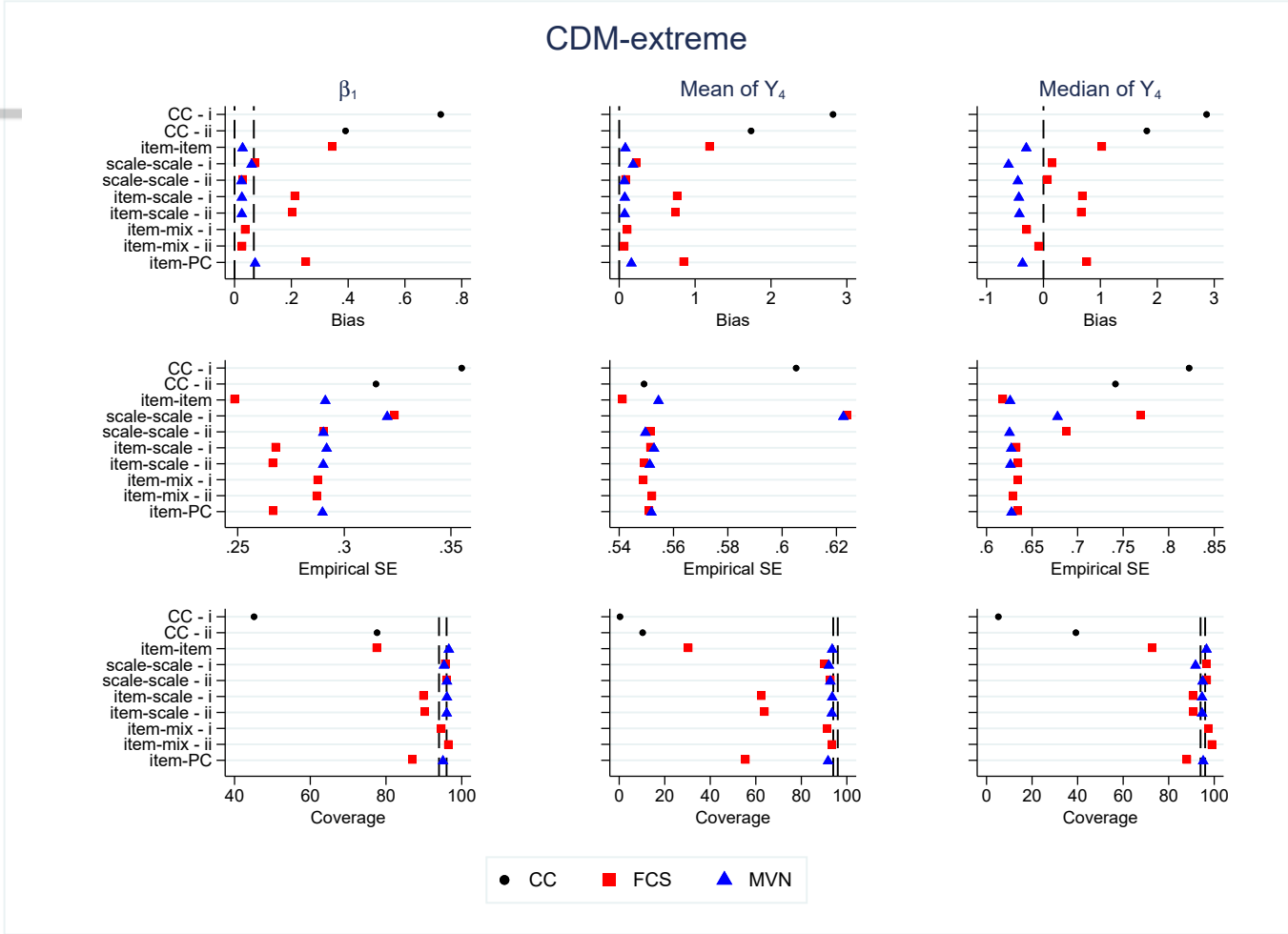

FIGURE 2 Bias, empirical standard error (SE) and coverage probability of the 95% confidence intervals for β_1, mean of Y_4 and median of Y_4, for the *CDM-extreme* scenario. Analysis was performed using either a complete cases analysis (CC), multiple imputation (MI) by fully conditional specification (FCS) or multivariate normal imputation (MVN). Five MI strategies were considered: *item-item*, item-level imputation using other items as auxiliary variables; *scale-scale*, scale-level imputation using other scale scores as auxiliary variables; *item-scale*, item-level imputation using scale scores as auxiliary variables; *item-mix*, item-level imputation using a mix of items and scales as auxiliary variables; *item-PC*, item-level imputation using principal component scores as auxiliary variables. Where applicable, scale scores calculated according to one of two rules (i and ii). Dashed horizontal reference lines are added to the bias figures at 0 and 10% of the true parameter value, and the coverage figures at 94% and 96%.

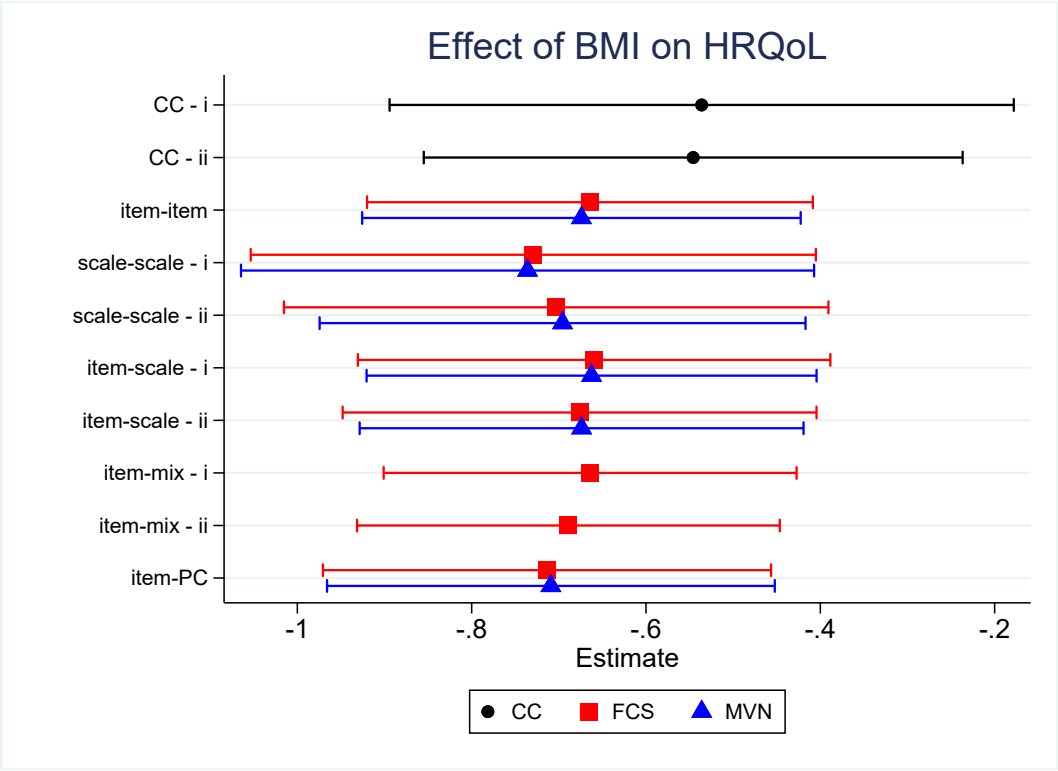

FIGURE 3 Estimate of the effect of body mass index (BMI) on health-related quality of life (HRQoL) from model (1), with 95% confidence intervals, from five different multiple imputation strategies implemented using either fully conditional specification (FCS) or multivariate normal imputation (MVN): *item-item*, item-level imputation using other items as auxiliary variables; *scale-scale*, scale-level imputation using other scale scores as auxiliary variables; *item-scale*, item-level imputation using scale scores as auxiliary variables; *item-mix*, item-level imputation using a mix of items and scales as auxiliary variables; *item-PC*, item-level imputation using principal component scores as auxiliary variables. Where applicable, scale scores calculated according to one of two rules (i and ii). A complete case analysis (CC) was performed for comparison.

How to cite this article: Mainzer R., Apajee J., Nguyen C., Carlin J.B., and Lee K.J. (2020), A comparison of multiple imputation strategies for handling missing data in composite scores with incomplete items: guidance for longitudinal studies, *Statistics in Medicine.*, *??*.