

PhD Completion seminar

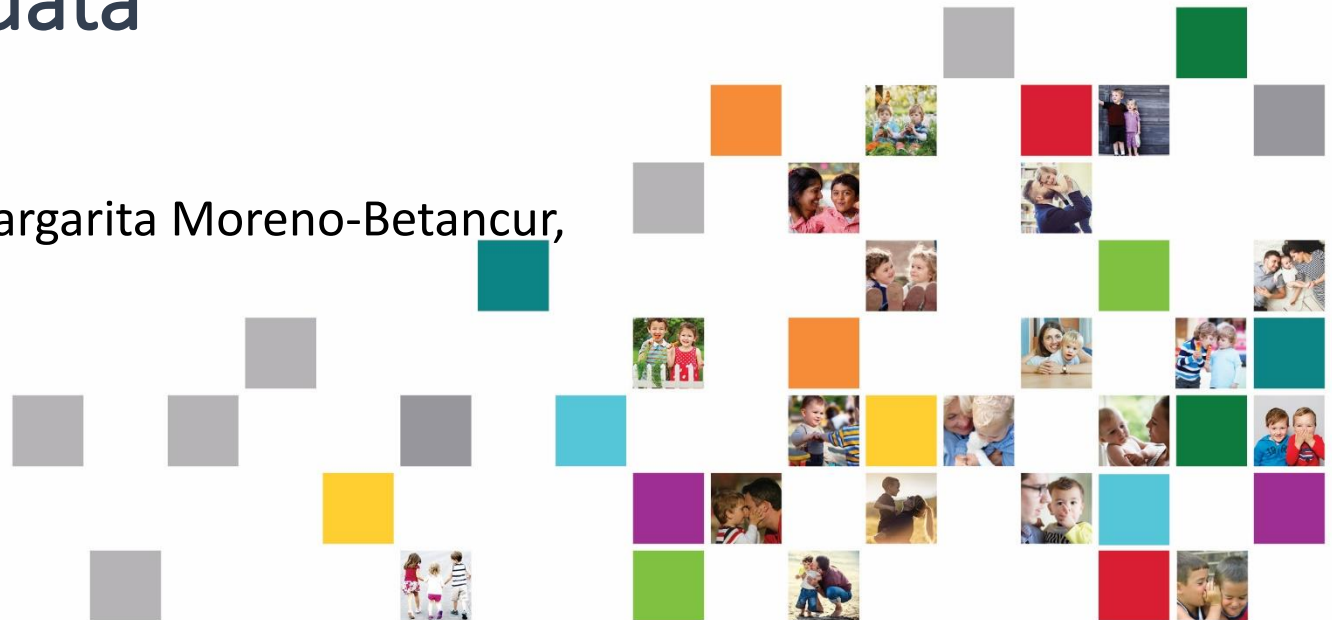
Evaluation of multiple imputation approaches for incomplete three-level data

Rushani Wijesuriya

Supervisors: Prof Katherine J. Lee, Dr. Margarita Moreno-Betancur,
Prof John B. Carlin, Dr. Anurika De Silva

Melbourne
Children's

Excellence in
clinical care,
research and
education



A bit of background



You can find more about my work :



[@rushwije](#)



<https://www.rwijesuriya.com/>



[@rush_099](#)

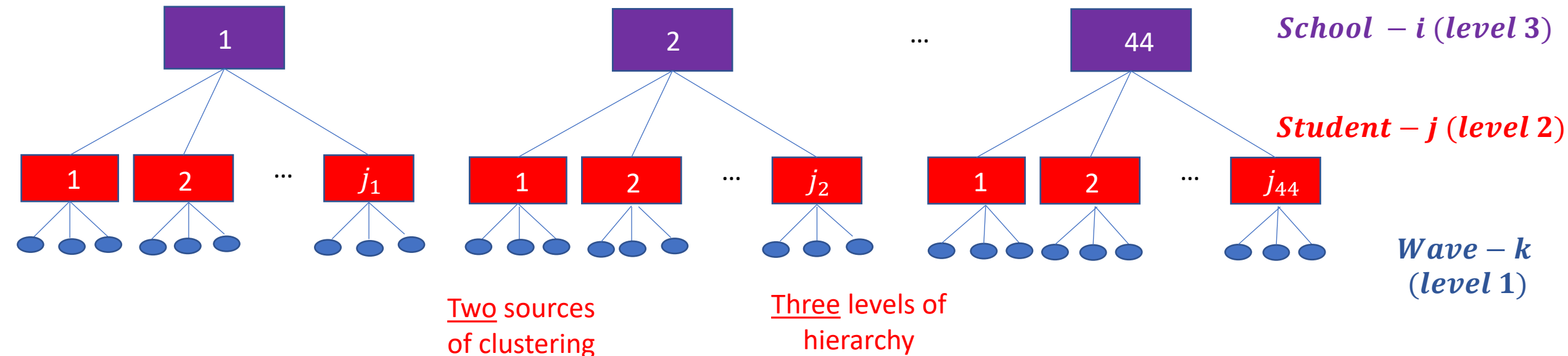
Overview of the seminar

- Background
- Overall research aim
- Research objective 1 (paper 1)
- Research objective 2 (paper 2)
- Research objective 3 (paper 3)
- Strength and limitations of the research
- Future work

Background

The Childhood to Adolescence Transition Study (CATS)

- A longitudinal cohort of 1239 students sampled from 44 schools at wave 1 (2012) and followed up annually



- Clustering of repeated measures within an individual and also clustering by school

Background

The Childhood to Adolescence Transition Study (CATS)

- Observations from units belonging to the same cluster are more likely to be correlated than observations from units belonging to different clusters
- These correlations need to be taken into account in statistical analyses
- With complete data, a widely used statistical modelling framework-linear mixed models (LMMs)

Background

Substantive research question

- The effect of early depressive symptoms on the academic performance of the students over time adjusted for confounders *accounting for clustering of individuals within schools and repeated measures within individuals*

Background

Target analysis models

- Different variations of LMMs
- Simplest used was a random intercepts model

$$\begin{aligned} NAPLAN_{ijk} = & \beta_0 + \beta_1 \times depression_{ij(k-1)} \\ & + \beta_2 \times wave \\ & + \beta_3 \times NAPLAN_{ij1} + \beta_4 \times sex_{ij} + \beta_5 \times SES_{ij1} + \beta_6 \times age_{ij1} \\ & + \boxed{b_{0i}} + \boxed{b_{0ij}} + \varepsilon_{ijk} \end{aligned} \quad \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

Where i denotes the i^{th} school, j denotes the j^{th} individual and k denotes the k^{th} wave

$$i = 1, \dots, 44$$

$$8 \leq j \leq 66$$

$$k = 3, 5, 7$$

Random intercept component for the school

$$b_{0i} \sim N(0, \sigma_{b_{0i}}^2)$$

Random intercept component for the individual

$$b_{0ij} \sim N(0, \sigma_{b_{0ij}}^2)$$

Background

The problem of missing data in the CATS

- In CATS most variables, specially those that are time-varying, contained missing data with the missingness percentage generally increasing over time

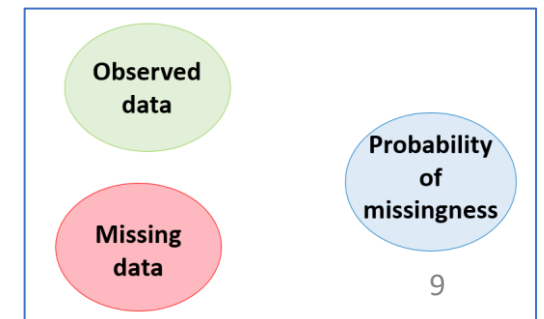
Background

Methods for handling missing data

- Available case analysis:

ID	Age	Sex	Wave	Dep
1	8	Male	1	0.4
1	8	Male	2	1.9
1	8	Male	3	0.2
2	7	Female	1	1.9
2	7	Female	2	✖
2	7	Female	3	2.9
3	10	Female	1	3.0
3	10	Female	2	✖
3	10	Female	3	✖

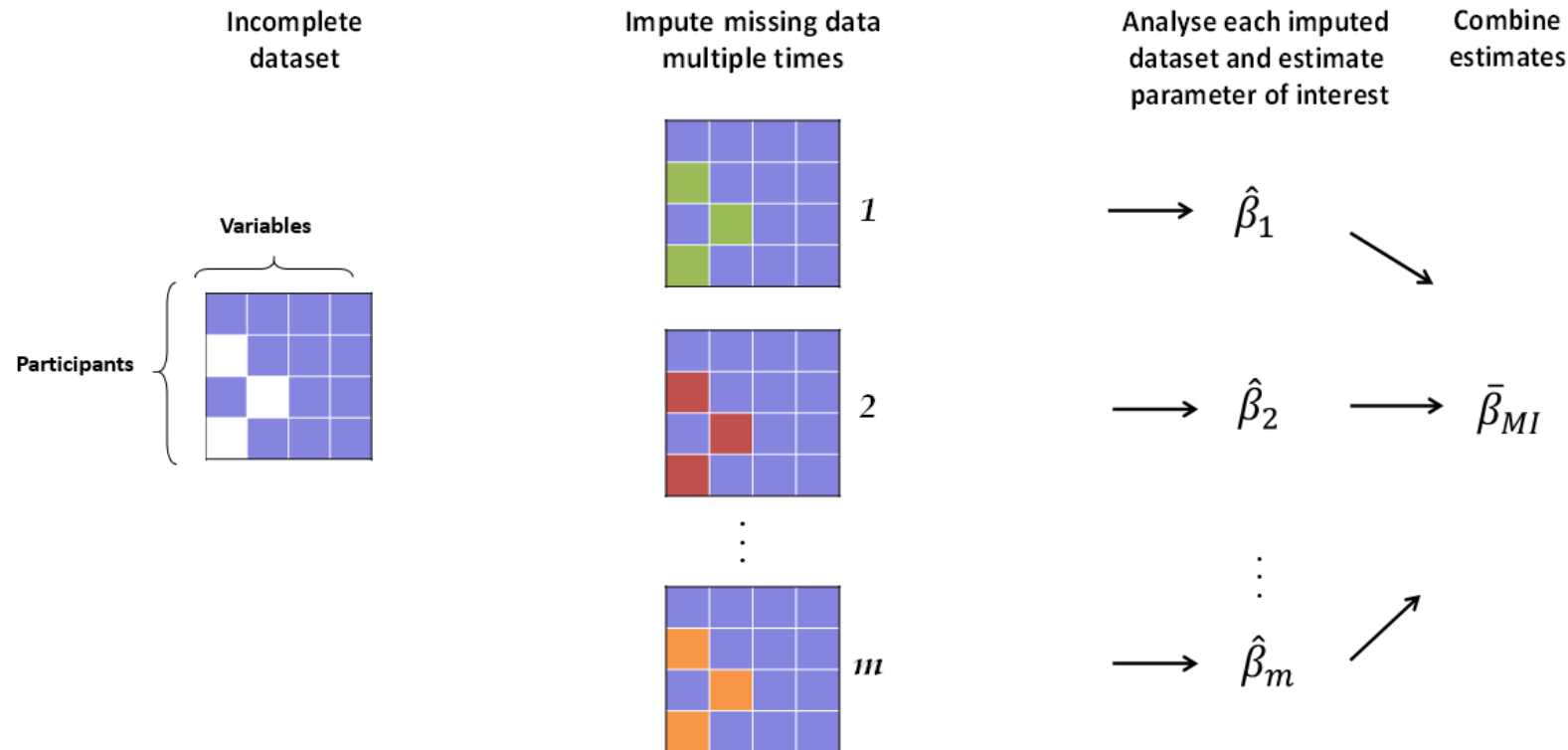
- Excludes waves with missing data
- Inferences can be biased
- Loss of information, leading to less precise estimates
- Valid under very strict assumptions about missing data - missing completely at random(MCAR)



Background

Methods for handling missing data

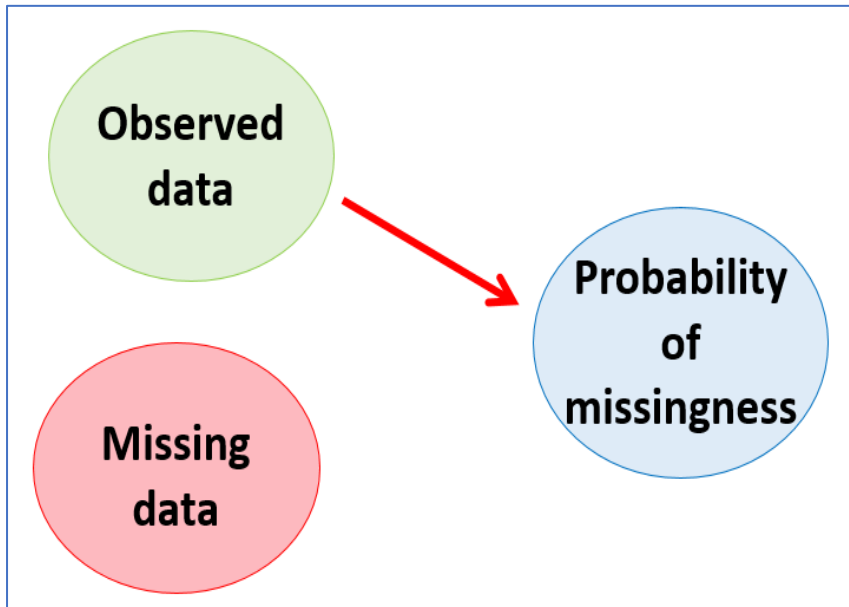
- Multiple imputation (MI)- a more principled approach



Background

Multiple imputation

- Assumptions:
 - Valid under more relaxed missing data assumptions –missing at random (MAR)



Background

Multiple imputation

- Implementation frameworks:
 - Joint modelling (JM)
 - Uses a joint imputation model to impute missing values
 - Imputes all variables with missing values simultaneously
 - Fully conditional specification(FCS)
 - Uses a series of univariate imputation models
 - Imputes variables sequentially

Background

Multiple imputation

- Model specification:
 - Inclusion of extra (auxiliary) variables to improve performance
 - Congeniality between the imputation and the analysis model**

Need to incorporate important features of the analysis (such as clustered structures, interactions, non-linearities etc) in the imputation model

Background

MI methods for handling missing non-clustered data: Single-level

- Conduct imputations assuming that the individual observations are independent
- Both JM and FCS implementations are available
- Most widely accessible in statistical software
 - Stata, R, SAS, SPSS

Background

MI methods for handling missing clustered data : Two-level

- Adaptations of single-level MI approaches
 - For cluster groups (such as schools): Dummy indicator (DI) approach
 - For repeated measures (at fixed intervals): Impute in wide format

ID	Age	Sex	Dep_1	Dep_2	Dep_3
1	8	Male	0.4	1.9	0.2
2	7	Female	1.9	.	2.9
3	10	Female	3.0	.	.
4	8	Male	.	2.6	.
5	10	Female	1.5	0.5	1.5



- MI approaches based on two-level models/LMMs

ID	Age	Sex	Wave	Dep
1	8	Male	1	0.4
1	8	Male	2	1.9
1	8	Male	3	0.2
2	7	Female	1	1.9
2	7	Female	2	.
2	7	Female	3	2.9
3	10	Female	1	3.0
3	10	Female	2	.
3	10	Female	3	.

- Both JM and FCS implementations available
 - R, SAS
- More recent than single-level implementations

Background

MI methods for handling missing clustered data : Three-level

- Adaptations of the single-level MI methods



- For cluster groups : Dummy indicator (DI) approach
- For repeated measures (at fixed intervals): Impute in wide format


ID	Age	Sex	Dep_1	Dep_2	Dep_3	school
1	8	Male	0.4	1.9	0.2	7
2	7	Female	1.9	.	2.9	25
3	10	Female	3.0	.	.	33
4	8	Male	.	2.6	.	10
5	10	Female	1.5	0.5	1.5	41


- JM-1L-DI-wide
- FCS-1L-DI-wide

Background

MI methods for handling missing clustered data : Three-level

- Adaptations of MI approaches based on two-level (RE) models

- 
- For cluster groups: DI approach
 - For repeated measures: Two-level MI approach (RE)

- 
- For cluster groups: Two-level MI approach (RE)
 - For repeated measures: Impute in wide format

ID	Age	Sex	Wave	Dep	School
1	8	Male	1	0.4	7
1	8	Male	2	1.9	7
1	8	Male	3	0.2	7
2	7	Female	1	1.9	25
2	7	Female	2	.	25
2	7	Female	3	2.9	25

- JM-2L-DI
- FCS-2L-DI

ID	Age	Sex	Dep_1	Dep_2	Dep_3	school
1	8	Male	0.4	1.9	0.2	7
2	7	Female	1.9	.	2.9	25
3	10	Female	3.0	.	.	33
4	8	Male	.	2.6	.	10
5	10	Female	1.5	0.5	1.5	33

- JM-2L-wide
- FCS-2L-wide

Background

MI methods for handling missing clustered data : Three-level

- MI approaches based on three-level (RE) models



- For cluster groups: RE
- For repeated measures: RE

ID	Age	Sex	Wave	Dep	school
1	8	Male	1	0.4	7
1	8	Male	2	1.9	7
1	8	Male	3	0.2	7
2	7	Female	1	1.9	25
2	7	Female	2	.	25
2	7	Female	3	2.9	25
3	10	Female	1	3.0	33
3	10	Female	2	.	33
3	10	Female	3	.	33

- Very recent
- Very limited implementations- mostly in stand-alone software or packages Ex: Blimp and R (miceadds package)
- Only FCS implementations are freely available
- JM-3L *
- FCS-3L

Background

Gaps in the literature

- Implementations of RE based methods for three-level data are not widely accessible
- Limited evaluations of ML approaches in the three-level context
- Lack of guidance in different settings

Research objectives

Overall aim

Evaluate MI approaches (pragmatic adaptations of the more widely available single- and two-level MI approaches and three-level MI approaches) for handling incomplete three-level data

with the aim of providing guidance on the use of these approaches in the context of a range of substantive analysis models.

Research objective 1

Paper 1

Evaluate MI approaches for handling incomplete three-level data in the setting of a random intercept (at both levels) analysis model using a simulation study

Wijesuriya et al. *BMC Medical Research Methodology* (2020) 20:207
<https://doi.org/10.1186/s12874-020-01079-8>


BMC Medical Research
Methodology

RESEARCH ARTICLE

Open Access

Evaluation of approaches for multiple imputation of three-level data



Rushani Wijesuriya^{1,2*} , Margarita Moreno-Betancur^{1,2}, John B. Carlin^{1,2} and Katherine J. Lee^{1,2}

[Link to paper](#)

Research objective 1

Recall: Substantive research question

- Effect of early depressive symptoms (at waves 2, 4 and 6) on the academic performance of the students (at waves 3, 5 and 7) as measured by NAPLAN numeracy scores

$$\begin{aligned} NAPLAN_{ijk} = & \beta_0 + \beta_1 \times depression_{ij(k-1)} \\ & + \beta_2 \times wave \\ & + \beta_3 \times NAPLAN_{ij1} + \beta_4 \times sex_{ij} + \beta_5 \times SES_{ij1} + \beta_6 \times age_{ij1} \\ & + b_{0i} + b_{0ij} + \varepsilon_{ijk} \end{aligned}$$

Research objective 1

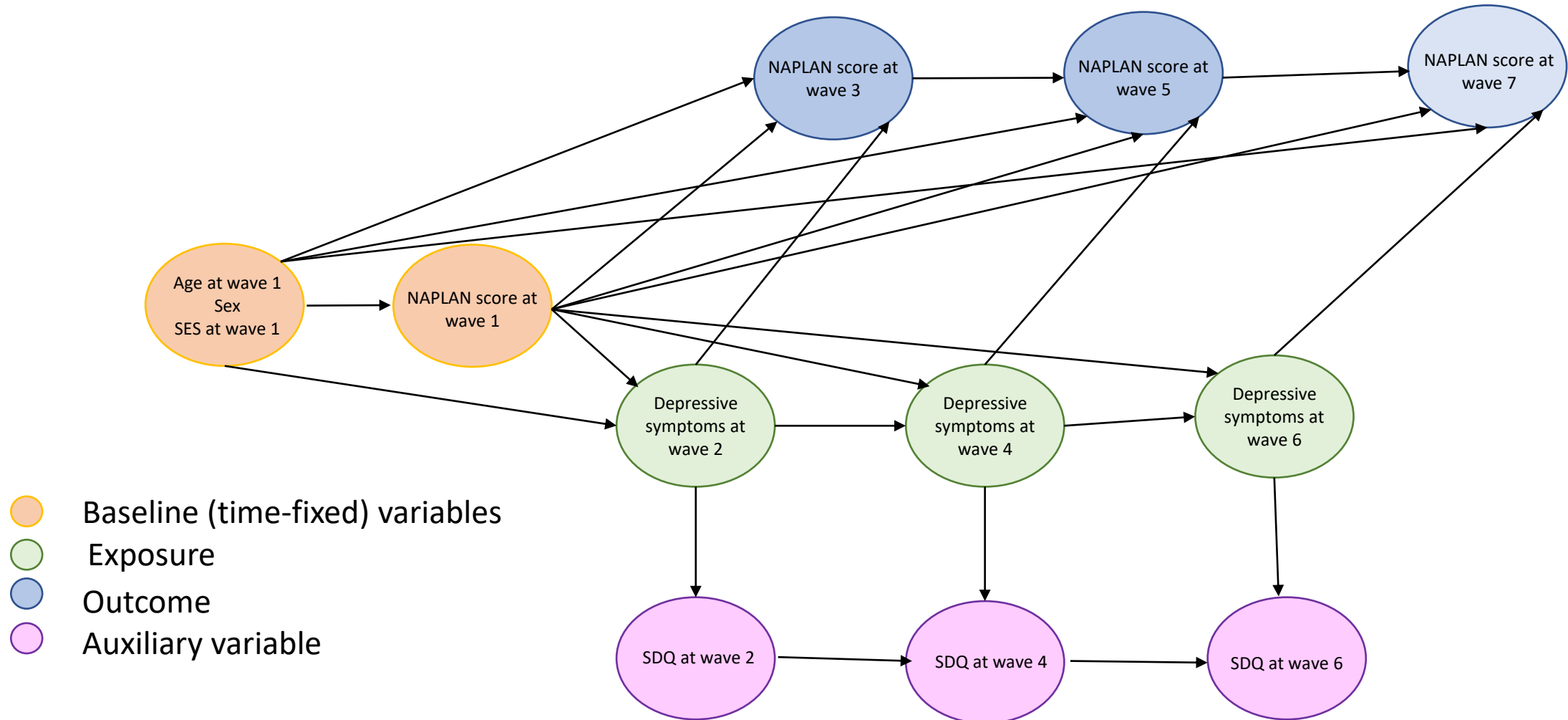
Simulation of complete data

- 1000 datasets were generated for each simulation condition
- 40 school clusters ($i = 1, \dots, 40$) were generated
- Each school cluster was populated
- Four different strengths of cluster correlations (intra-cluster correlation) at the school-level and the individual level were considered

	ICC	
	level 3 (within school)	level 2 (within individual)
High-high	0.15	0.5
High-low	0.15	0.2
Low-high	0.05	0.5
Low-low	0.05	0.2

Research objective 1

Simulation of complete data



Research objective 1

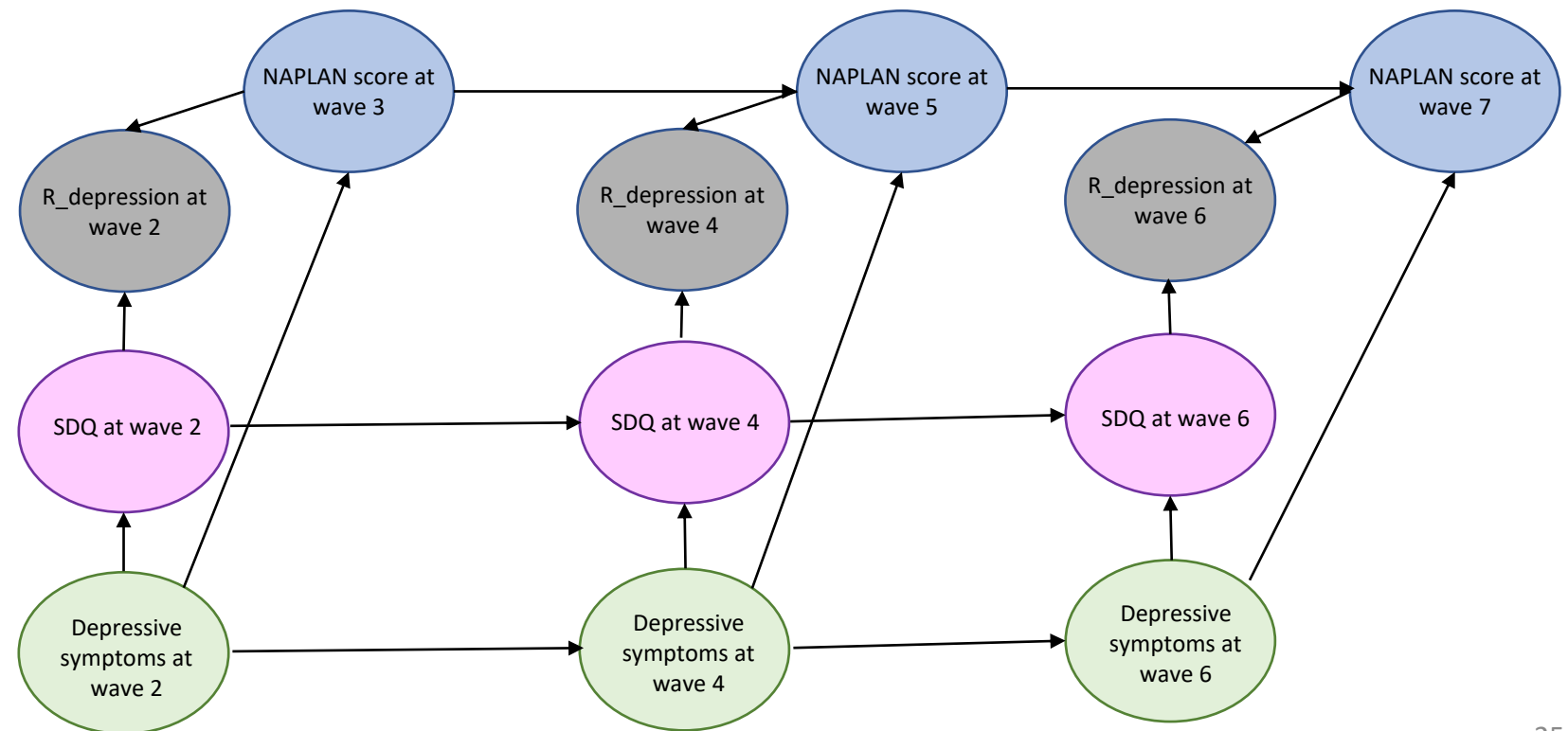
Generation of missing data

10% ← Depressive symptoms at wave 2 → 20%
15% ← Depressive symptoms at wave 4 → 30%
20% ← Depressive symptoms at wave 6 → 40%

MCAR
Missing
values
assigned
completely
at random

MAR

Similar to CATS (MAR-CATS)
Strong (MAR-inflated)



Research objective 1

Parameters of interest and performance measures

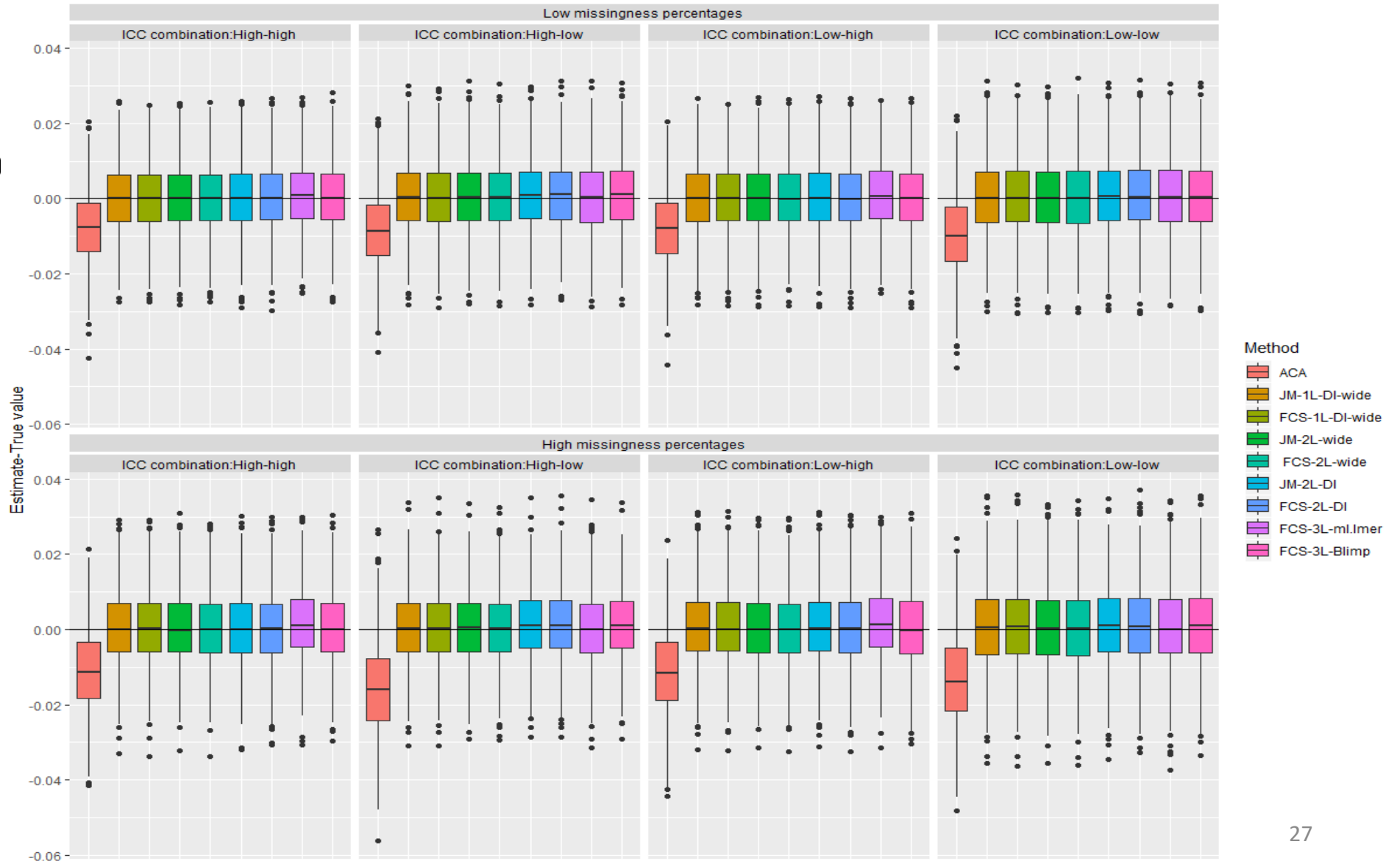
$$\begin{aligned} NAPLAN_{ijk} = & \beta_0 + \beta_1 \times depression_{ij(k-1)} \\ & + \beta_2 \times wave \\ & + \beta_3 \times NAPLAN_{ij1} + \beta_4 \times sex_{ij} + \beta_5 \times SES_{ij1} + \beta_6 \times age_{ij1} \\ & + b_{0i} + b_{0ij} + \varepsilon_{ijk} \\ & b_{0i} \sim N(0, \sigma_{b_{0i}}^2) \quad b_{0ij} \sim N(0, \sigma_{b_{0ij}}^2) \quad \varepsilon_{ijk} \sim N(0, \sigma_{\varepsilon}^2) \end{aligned}$$

- Deviations from the true value
- Average bias
- Empirical standard error
- Model-based standard error
- Coverage

Research objective 1

Results

- $\beta_1 = (-0.025)$

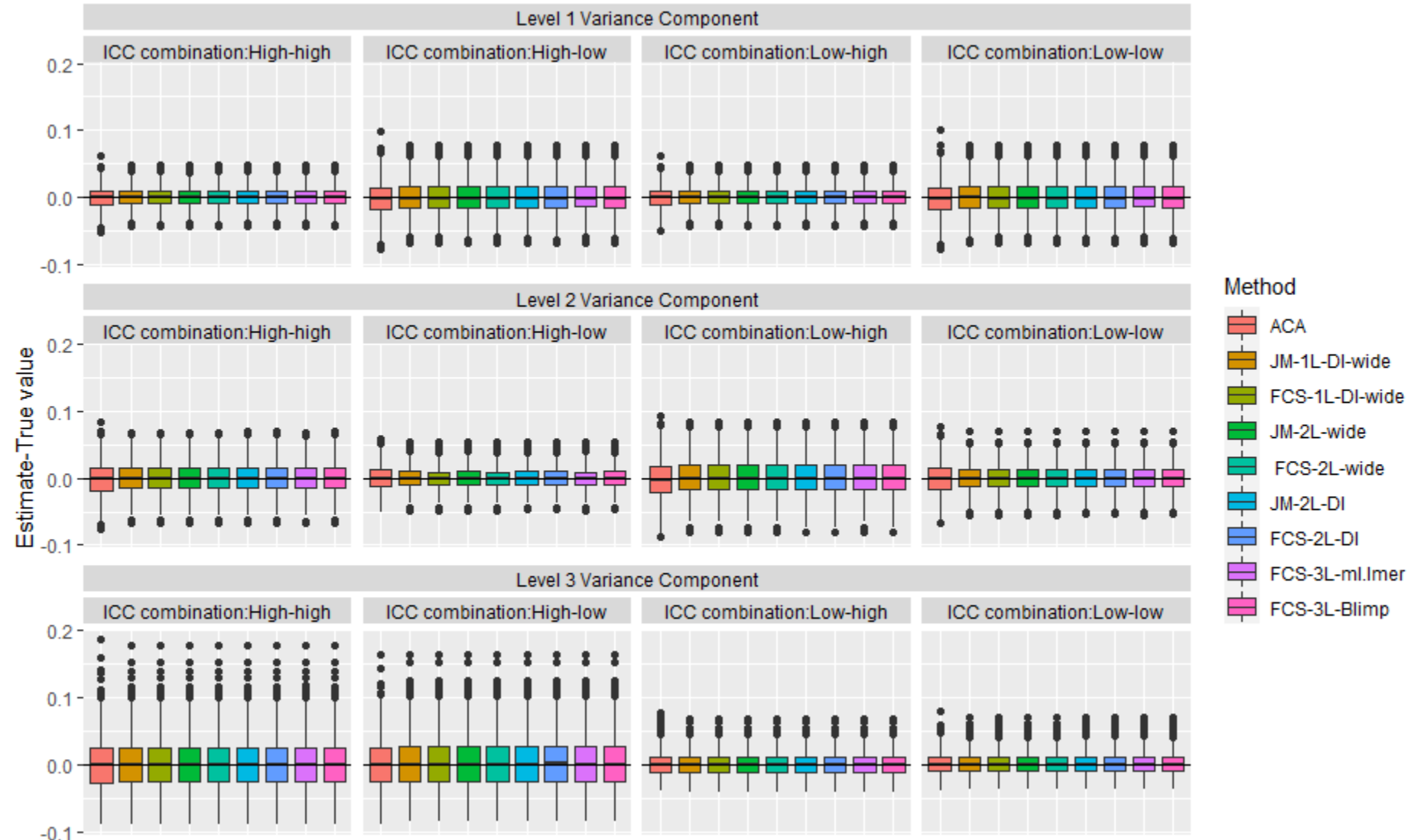


Results shown only for MAR-CATS scenario as the comparative performance of approaches was similar under MAR-inflated scenario

Research objective 1

Results

Variance
Components
(for the low
missingness
Percentages)



Research objective 1

Recommendations

- Adaptations of the single-and two-level MI approaches can be used to adequately handle incomplete three-level data in the context of a random intercept substantive analysis model
- Approaches which use the DI extension should be used with caution as it has been shown to produce biased parameter estimates in certain scenarios
- In the presence of longitudinal data measured at irregular time intervals, three-level imputation approaches will be required

Research objective 2

Paper 2 [under review after revisions at Biometrical journal]

Evaluate ML approaches for handling incomplete three-level data in the presence of **interactions and non-linear terms**

[Link to pre-print](#)

Research objective 2

Recall: Substantive research question

- The effect of early depressive symptoms on the academic performance of the students over time adjusted for confounders *accounting for clustering of individuals within schools and repeated measures within individuals*
- Three random intercept LMM analysis models that are common in longitudinal data settings were considered:

Research objective 2

Target analysis models

1. An interaction between the time-varying exposure and time

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave + \beta_3 \times depression_{ij(k-1)} \times wave + ** + b_{0i} + b_{0ij} + \varepsilon_{ijk}$$

2. An interaction between the time-varying exposure and a time-fixed baseline variable

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave + \beta_3 \times depression_{ij(k-1)} \times SES_{ij1} + ** + b_{0i} + b_{0ij} + \varepsilon_{ijk}$$

3. A quadratic effect of the time-varying exposure

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave + \beta_3 \times depression_{ij(k-1)}^2 + ** + b_{0i} + b_{0ij} + \varepsilon_{ijk}$$

**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

Research objective 2

Accommodating interactions and non-linear terms within MI methods for handling three-level data

To ensure congeniality between the imputation and analysis models: need to incorporate the three-level structure *and* the non-linear or interaction terms

Approaches to include interactions or non-linear terms include:

- Just another variable (JAV) approach
- Passive imputation
- Substantive model compatible (SMC) approaches

Research objective 2

Accommodating interactions and non-linear terms within MI methods for handling three-level data

Adaptations of the single-level MI methods

- Cluster groups : DI approach
- Repeated measures: Impute in wide format

Adaptations of MI approaches based on two-level (RE) models

- Cluster groups: Two-level MI approach (RE)
- Repeated measures: Impute in wide format

Adaptations of MI approaches based on two-level (RE) models

- Cluster groups: DI approach
- Repeated measures: Two-level MI approach (RE) (imputed in long format)

MI approaches based on three-level (RE) models (repeated measures imputed in long format)

- Cluster groups: RE
- Repeated measures: RE

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

SMC-JM-3L

Research objective 2

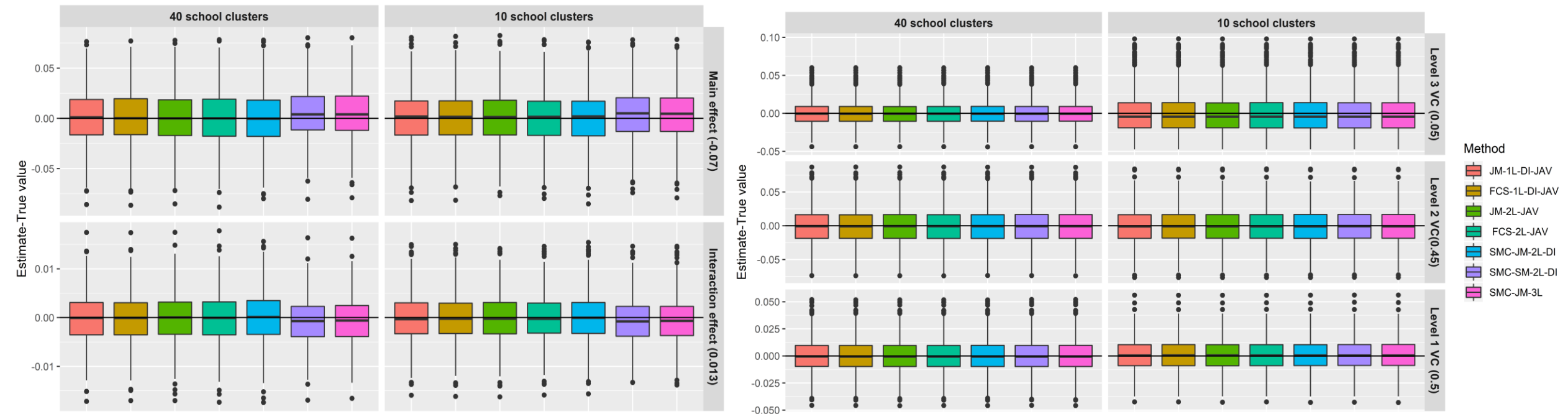
Simulation of complete data

- Simulation of complete data similar to research objective 1
- 1000 datasets were simulated
- Two different numbers of higher-level clusters considered : 40 school clusters and 10 school clusters
- Missing values generated in:
 - The exposure(15%, 20% and 30% of the depressive symptom scores at waves 2,4, and 6 respectively) according to a MAR mechanism (MAR-CATS, MAR-inflated)
 - a time-fixed confounder (10 % of Socio-Economic Status) according to a MCAR mechanism

Research objective 2

Results - Analysis model 1

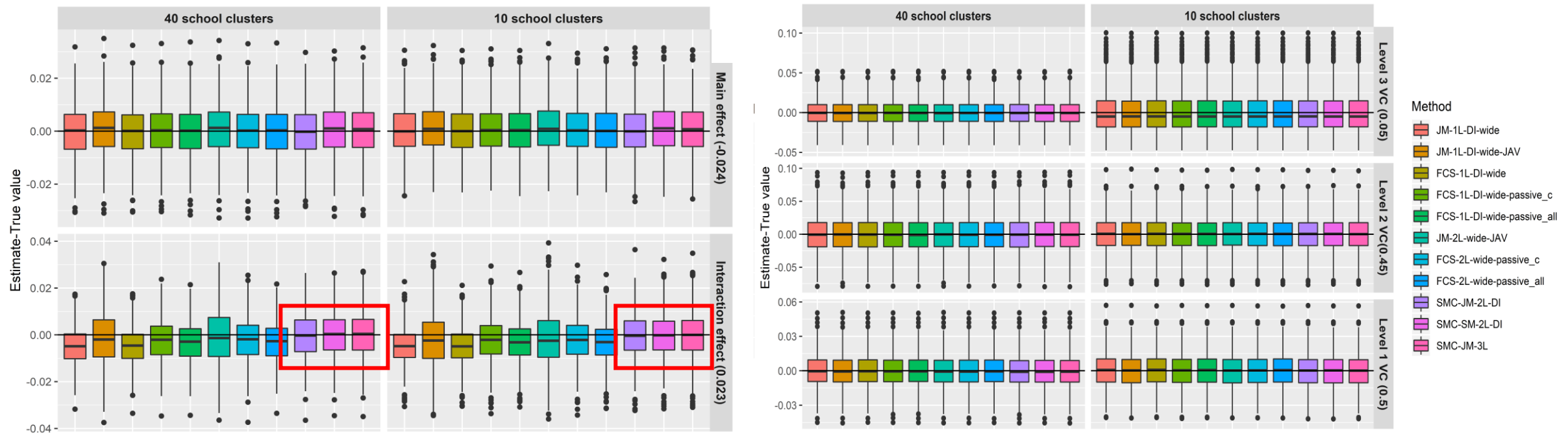
Interaction between the time-varying exposure and time



Research objective 2

Results-Analysis model 2

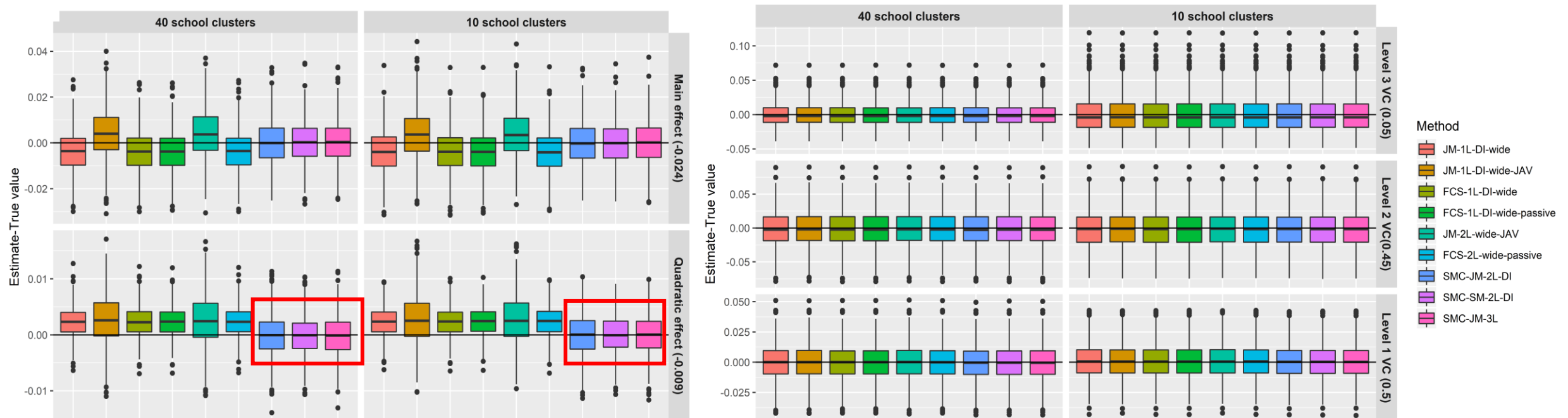
Interaction between the time-varying exposure and a time-fixed baseline variable



Research objective 2

Results - Analysis model 3

Quadratic effect of the time-varying exposure



Research objective 2

Recommendations

- With an analysis model where there is an interaction with time, all of the approaches (including the single-level and two-level adaptations) seem to be appropriate
- When the analysis model involves an interaction between the time-varying exposure and an incomplete time-fixed confounder or quadratic effects the three-level SMC approach is recommended

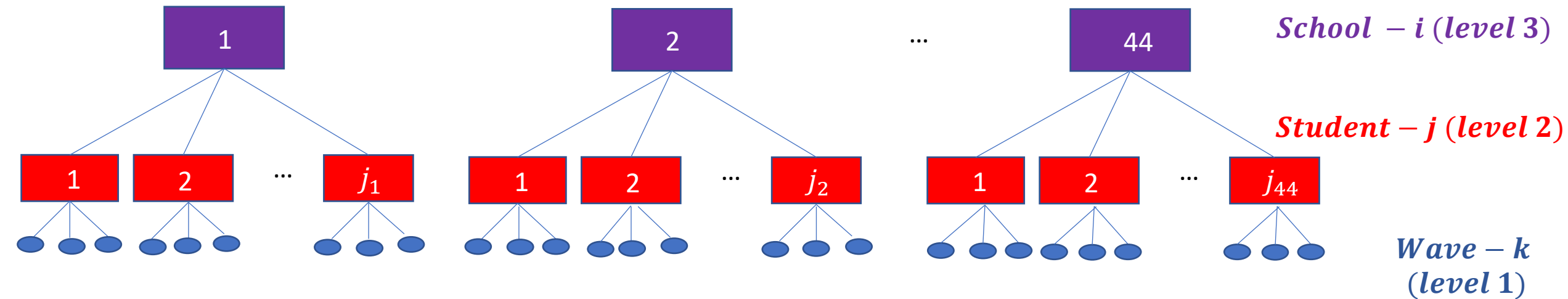
Research objective 3

Paper 3 [Manuscript in preparation]

Evaluate ML approaches for handling incomplete three-level data with time-varying higher-level cluster memberships

Research objective 3

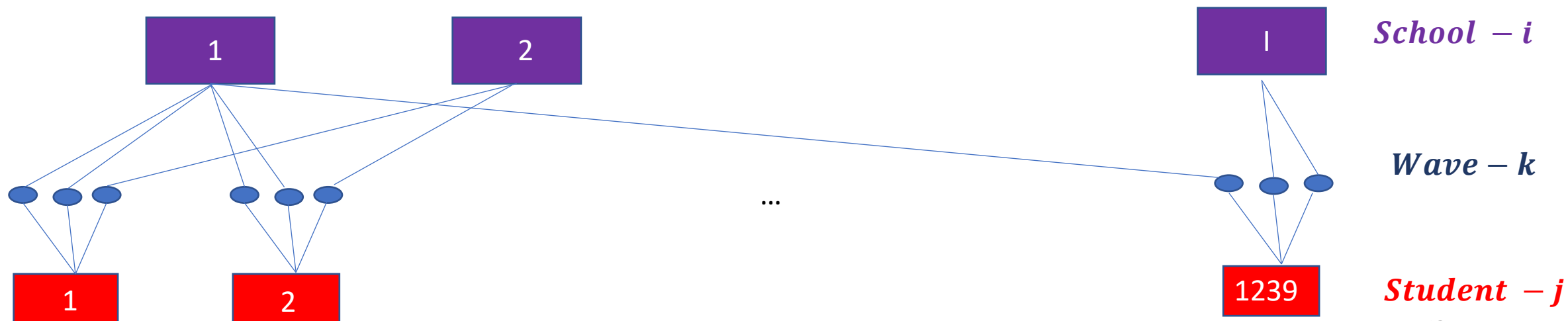
The Childhood to Adolescence Transition Study (CATS)



Research objective 3

The Childhood to Adolescence Transition Study (CATS)

- Children moved schools over time
- Resulted in a cross-classified structure where the repeated measures are clustered within individuals but the individuals are no longer clustered within the same higher-level cluster throughout the study period



Research objective 3

Recall: Substantive research question

The effect of early depressive symptoms on the academic performance of the students over time adjusted for confounders *accounting for clustering of individuals within schools and repeated measures within individuals*

In this work the focus was on:

estimating the effect of depressive symptoms (at waves 1, 2 and 3, exposure) on the subsequent academic performance of the students (at waves 2, 3 and 4, outcome) as measured by a numeracy rating provided by the school teacher

Research objective 3

Target analysis model

A cross-classified random effects model (CCREM)

The **random intercept at the school level** was time-varying to allow for the effect of participant's school (highest level group membership) to vary from wave to wave

$$\begin{aligned} teacher_score_{ijk} = & \beta_0 + \beta_1 \times depression_{i'j(k-1)} \\ & + \beta_2 \times wave \\ & + \beta_3 \times teacher_score_{j1} + \beta_4 \times sex_j + \beta_5 \times SES_{j1} + \beta_6 \times age_{j1} \\ & + \alpha_i + \gamma_j + \varepsilon_{ijk} \end{aligned}$$

Where i ($i \in 1, \dots, 181$) denote the school the individual j ($j = 1, \dots, 1168$) attended at wave k ($k = 2, 3, 4$) and i' indicates the school cluster membership of the student at the wave ($k-1$)

Research objective 3

Accommodating time-varying cluster memberships within MI methods for handling three-level data

To ensure congeniality between the imputation and analysis models: need to incorporate the cross-classified structure in the imputation model

Some ad-hoc approaches include:

- First-cluster approach
- Common-cluster approach



Can be implemented within any approach for handling three-level data

- JM-1L-DI-wide_f, FCS-1L-DI-wide_f, FCS-3L-Blimp_f
- JM-1L-DI-wide_c, FCS-1L-DI-wide_c, FCS-3L-Blimp_c

Research objective 3

Accommodating time-varying cluster memberships within MI methods for handling three-level data

Adaptations of the single-level MI methods

- Cluster groups : DI approach
- Repeated measures: Impute in wide format

Adaptations of MI approaches based on two-level (RE) models

- Cluster groups: Two-level MI approach (RE)
- Repeated measures: Impute in wide format

Adaptations of MI approaches based on two-level (RE) models

- Cluster groups: DI approach
- Repeated measures: Two-level MI approach (RE) (imputed in long format)

MI approaches based on three-level (RE) models (repeated measures imputed in long format)

- Cluster groups: RE
- Repeated measures: RE

Accommodating time-varying cluster memberships

- Not possible within JM framework as the repeated measures in wide format are imputed simultaneously
- Within FCS, include the cluster membership at the current wave in each univariate imputation model specified for each incomplete repeated measure

DIs to represent the time-varying cluster memberships in long format

- JM-using a multivariate CCREM^{*}
- FCS- series of univariate CCREMs

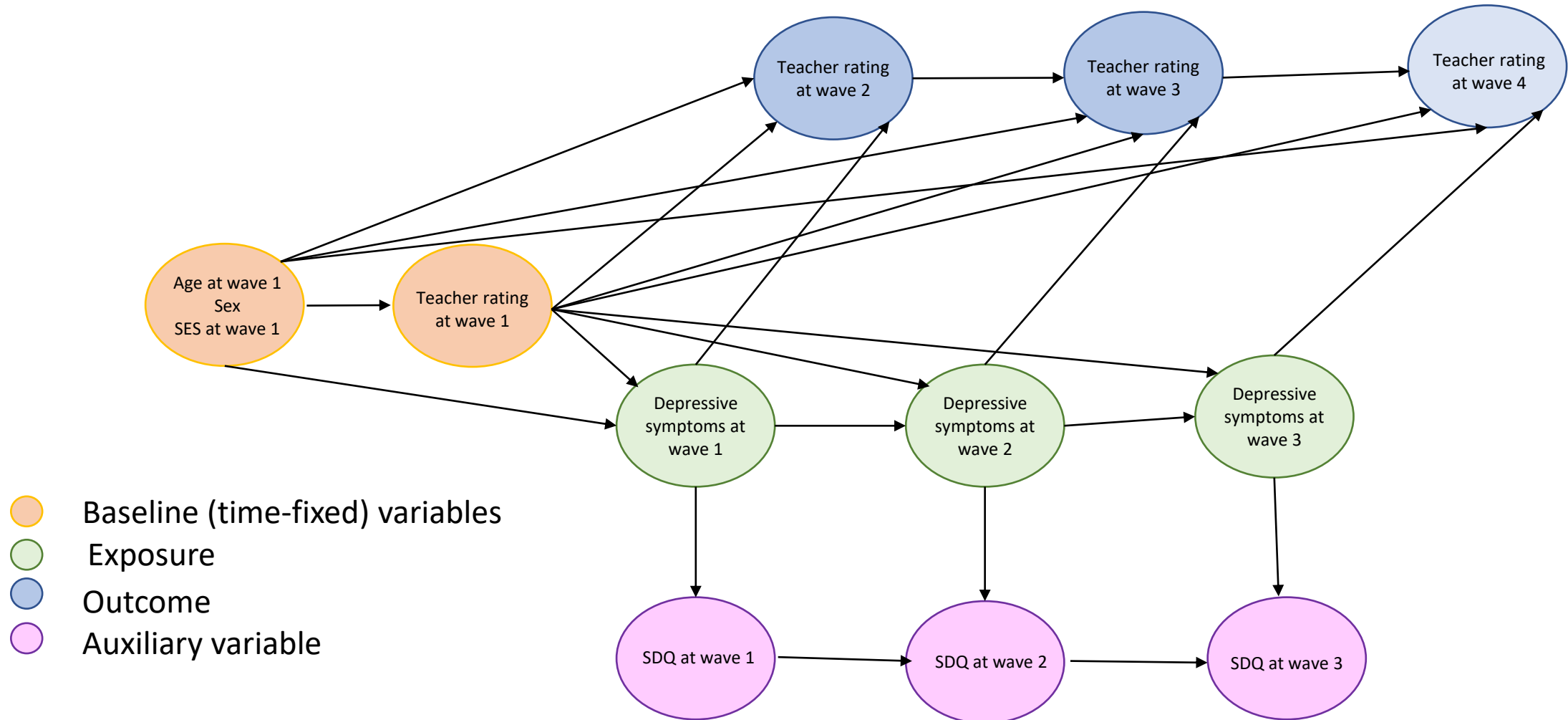
Research objective 3

Simulation of complete data

- 40 school clusters (at wave 1) were generated, and were populated with varying number of students (totalling to 1200)
- Baseline variables were first generated for each student
- New school clusters (50 and 10) were added at waves 2,3 and 4
- 5% of students at each wave were then randomly selected to be moved to these newly generated schools, with equal numbers of students being assigned to each new school
- Time-varying variables were finally generated
- Four different strengths of cluster correlations at the school-level and the individual level were also considered (similar to [paper 1](#))

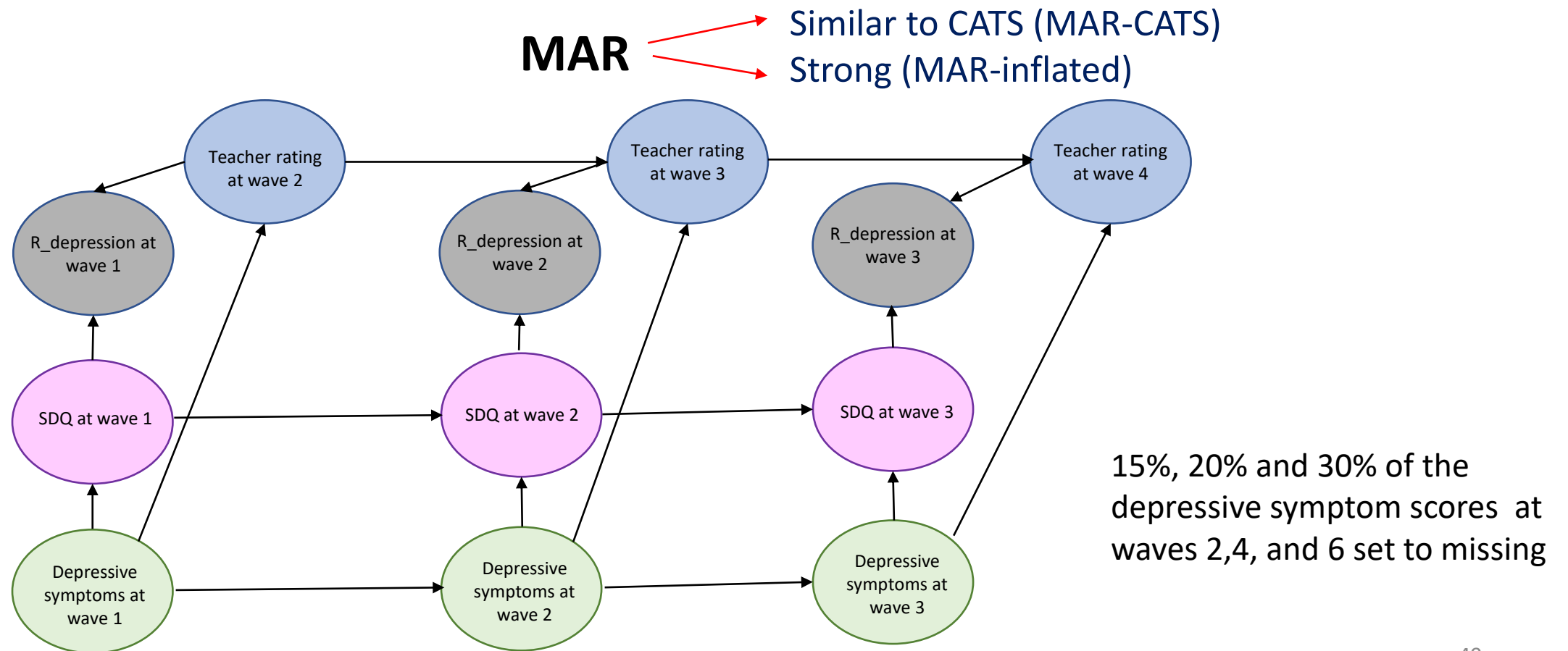
Research objective 3

Simulation of complete data



Research objective 3

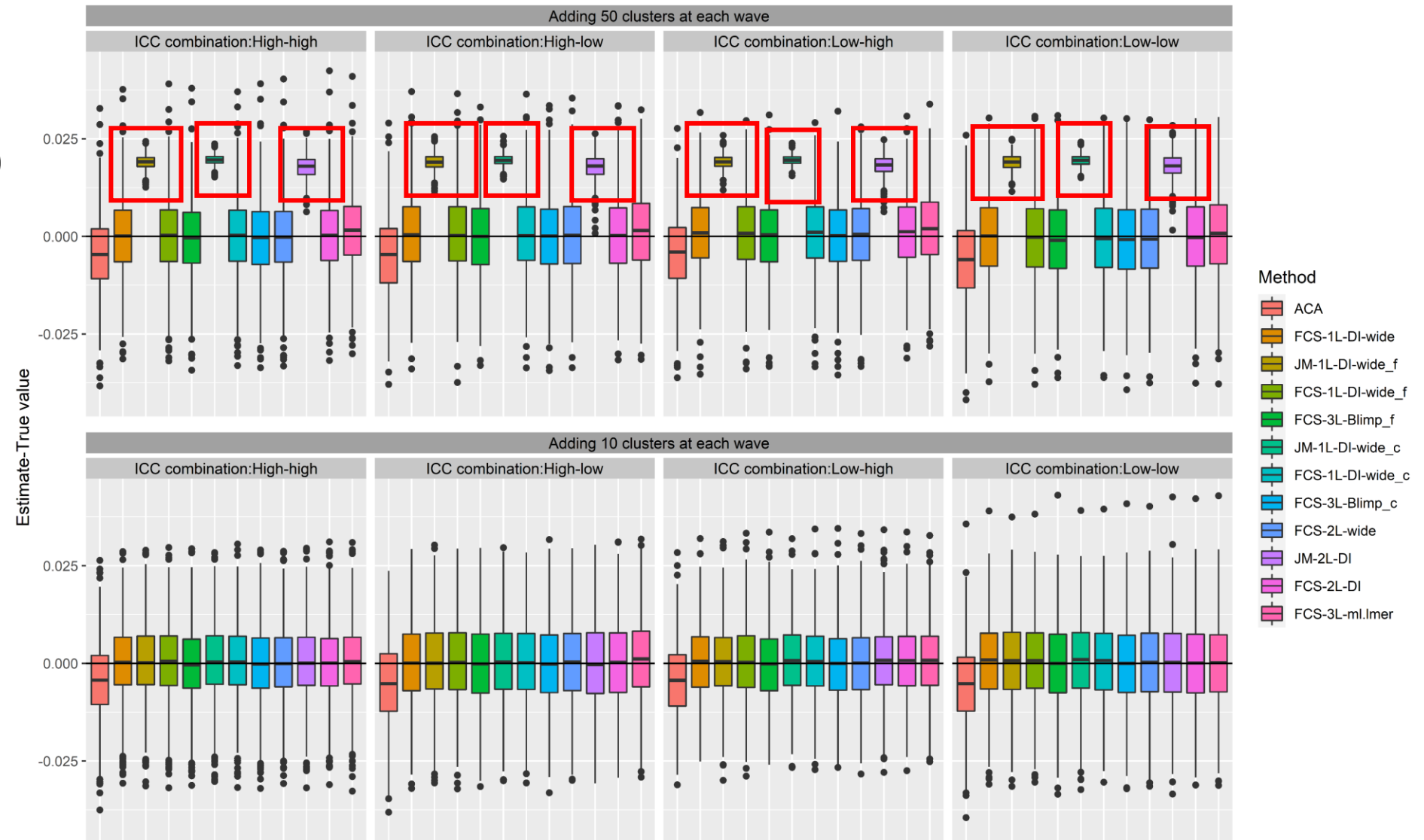
Generation of missing data



Research objective 3

Results

$$\beta_1 = (-0.02)$$

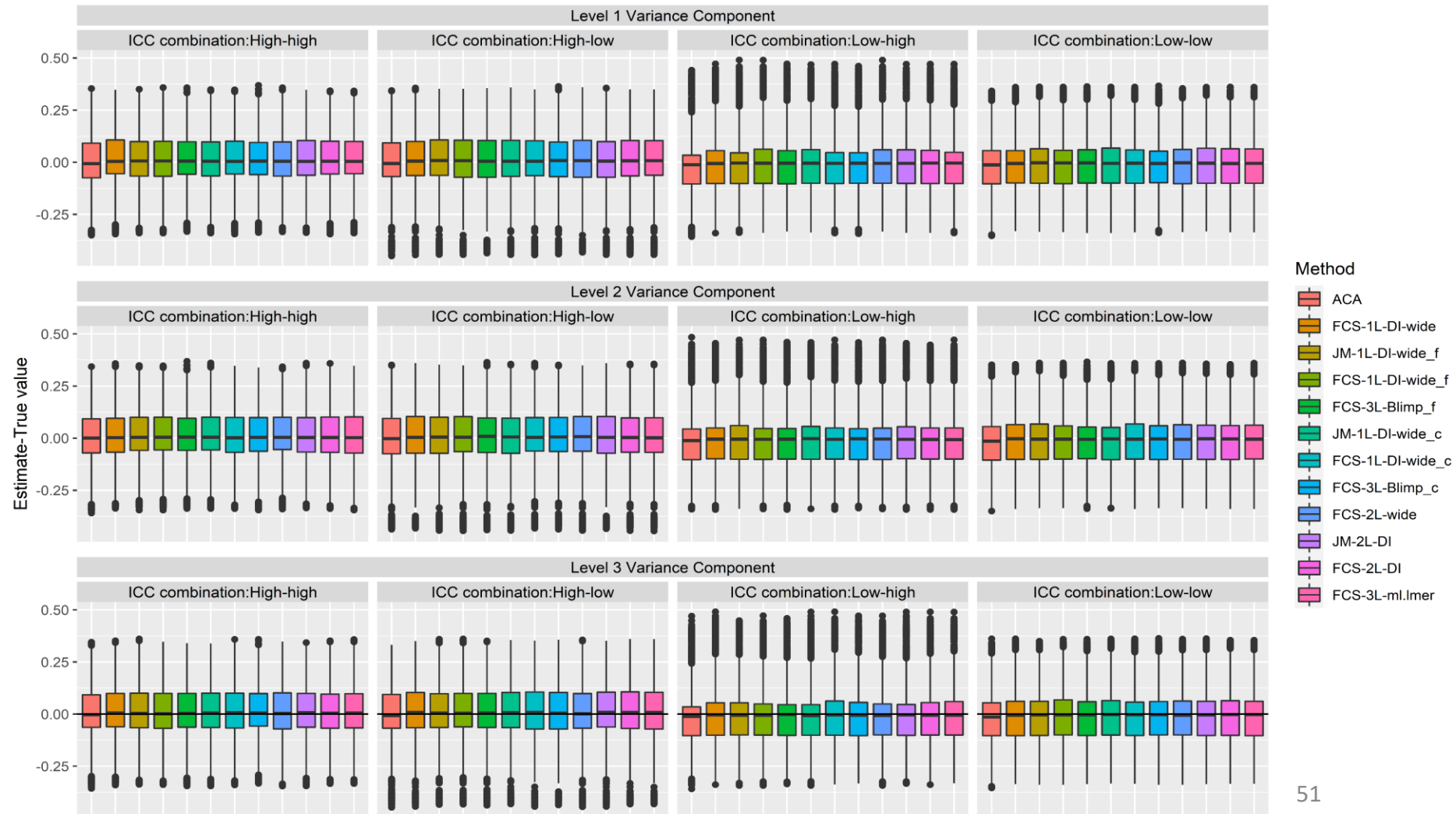


Research objective 3

Results

Variance components

Adding 50 clusters at each wave



Research objective 3

Recommendations

- The extensions of the single-level and two-level FCS approaches or the three-level FCS approach can be used to handle incomplete three-level cross classified data
- The three-level FCS approach may need to be used in settings with irregularly measured time points
- Use extensions of the JM approaches with caution as the large number of DIs in these approaches can lead to biased estimates of the regression coefficient estimates and the variance components

Strengths and limitations

Strengths

- First to provide an overview and a systematic comparison of all available MI approaches for three-level data under various contexts
- Simulations based on a real cohort study

Limitations

- Limited set of scenarios
- Missingness imposed in continuous variables only
- Simulation design may favour some imputation approaches

Future work

- More complex random effects substantive analysis models (both hierarchical and cross-classified)
- MNAR missingness scenarios
- Comparative performance of the approaches with mixtures of incomplete continuous and categorical variables
- Missingness at the cluster level