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

Rushani Wijesuriya

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

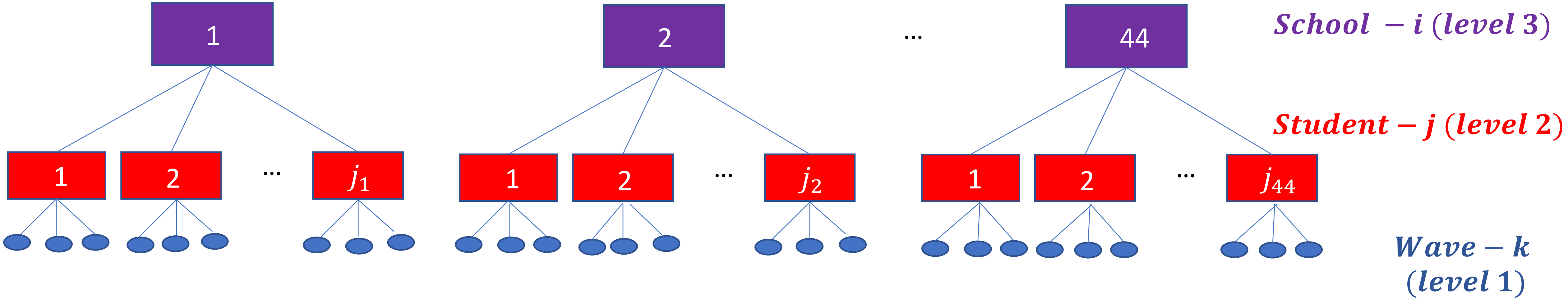
ISCB 2021

Overview

- Background: the motivating example
- MI methods for accommodating time-varying cluster memberships
- Simulation study
- Simulation results
- Case study illustration
- Conclusions

Background: the motivating example

The Childhood to Adolescence Transition Study (CATS)

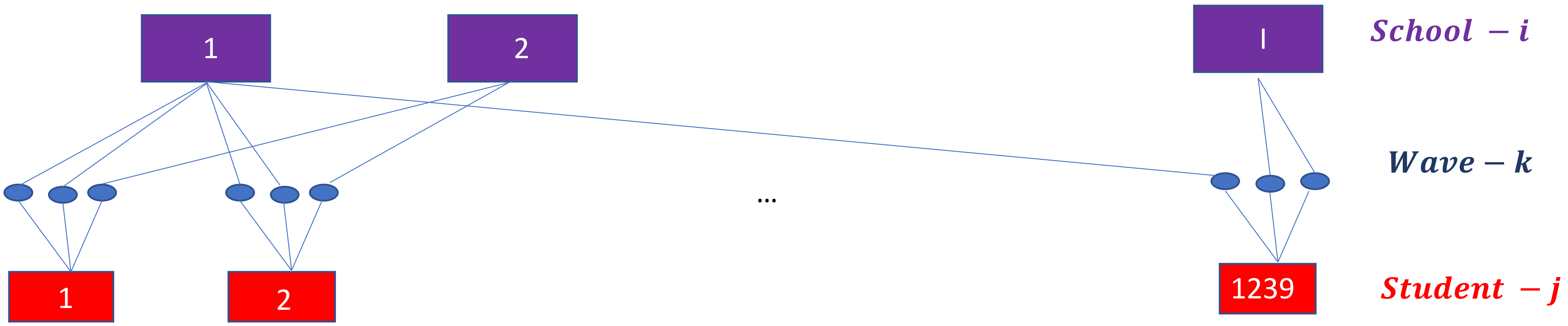


- Repeated measures of students nested within schools

Background: the motivating example

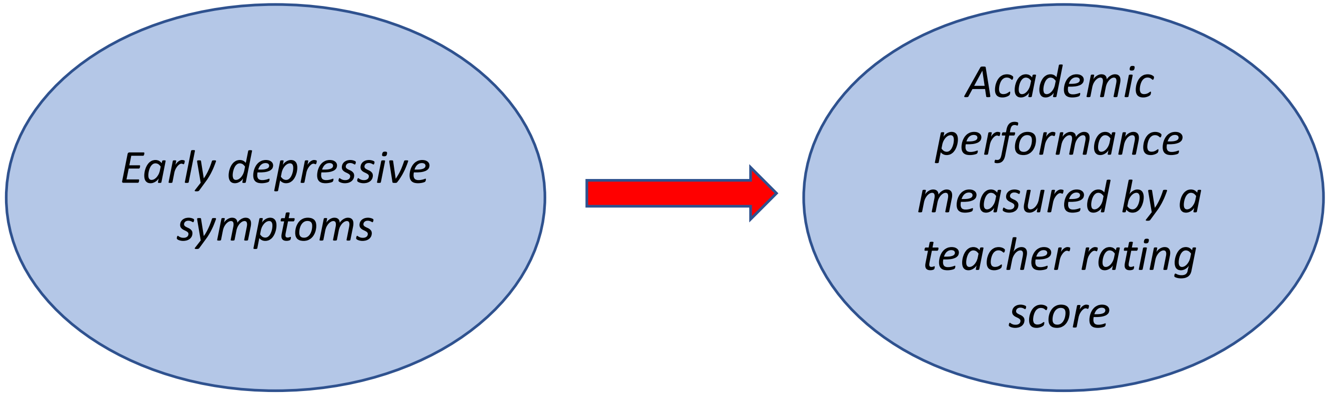
The Childhood to Adolescence Transition Study (CATS)

- Additional complexity: children moved schools over time
- A cross-classified structure : the repeated measures are clustered within individuals but the individuals are no longer clustered within the same higher-level cluster throughout the study period



Background: the motivating example

The substantive research question



accounting for clustering of individuals within schools and repeated measures within individuals

Background : the motivating example

Target analysis model

A cross-classified random effects model (CCREM)

A time-varying random intercept at the school level to allow for the effect of participant's school to vary from wave to wave

$$
\begin{aligned}
teacher\_score_{jk} = \beta_0 + \beta_1 \times depression_{j(k-1)} \\
+\beta_2 \times wave \\
+\beta_3 \times teacher\_score_{j1} + \beta_4 \times sex_j + \beta_5 \times SES_{j1} + \beta_6 \times age_{j1} \\
+\alpha_{i(j,k)} + \gamma_j + \varepsilon_{jk}
\end{aligned}
$$

Where i (j,k) denote the school the individual j (j=1, …, 1168) attended at wave k (k=2,3,4)

Background: the motivating example

The problem of missing data in CATS

- In CATS missing data were observed in all the time-varying variables
- MI is a popular approach for handling incomplete data
- Most common MI frameworks: Joint modelling (JM) and fully conditional specification (FCS)
- Under both these frameworks : need to ensure congeniality between the imputation and analysis model for valid results
- Appropriately tailoring the imputation model to include important features of the substantive analysis

Need to incorporate the two sources of clustering and the time-varying cluster memberships in the imputation model

Methods

MI methods for handling three-level data

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

Methods

MI methods for handling three-level data

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

- For cluster groups: Two-level MI approach (RE)
- For repeated measures: 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	33

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

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

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

RE: Random effects

Methods

MI methods for handling three-level data

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

- Two FCS implementations are freely available
  - FCS-3L-ml.lmer (R-miceadds)
  - FCS-3L-Blimp

RE: Random effects

Methods

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

Two ways:

1. Ignore either by using the first cluster or the most common cluster

   first cluster approach : JM-1L-DI-WIDE_f, FCS-1L-DI-WIDE_f, FCS-3L-Blimp_f

   common cluster approach: JM-1L-DI-WIDE_c, FCS-1L-DI-WIDE_c, FCS-3L-Blimp_c

2. Accommodate within the MI approach

Methods

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

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

- FCS-1L-DI-wide
- FCS-2L-wide

Vary cluster memberships in long format

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

- Using CCREM as the imputation model

- FCS-3L-ml.lmer

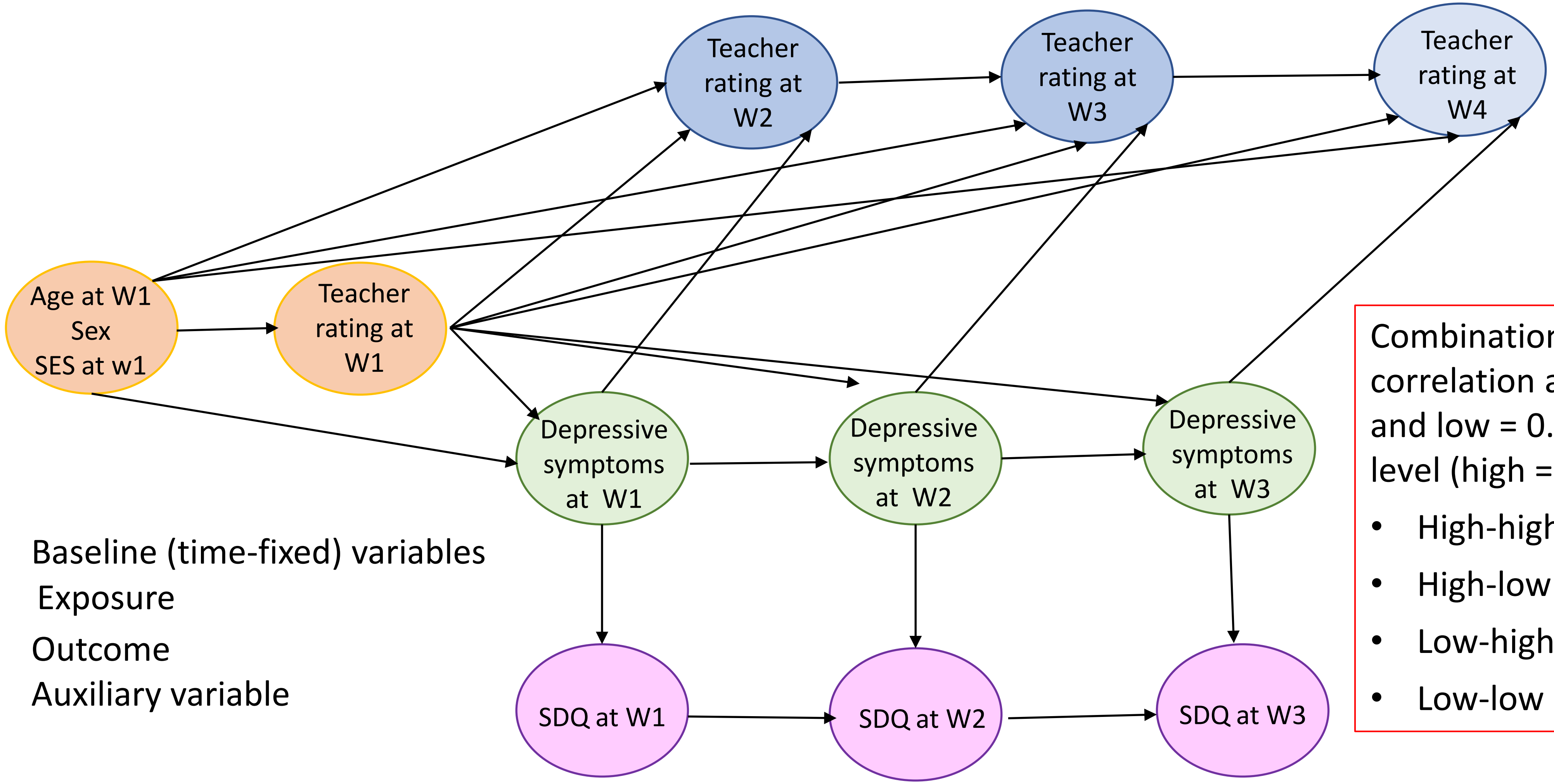
RE: Random effects, DI:

Simulation study

Simulation of complete data

To mimic the cross-classified structure:

- New school clusters (50 and 10) were added at waves 2,3 and 4
- 5% of students at each wave were selected randomly to move to these schools, with equal numbers of students assigned to each



Baseline (time-fixed) variables
Exposure
Outcome
Auxiliary variable

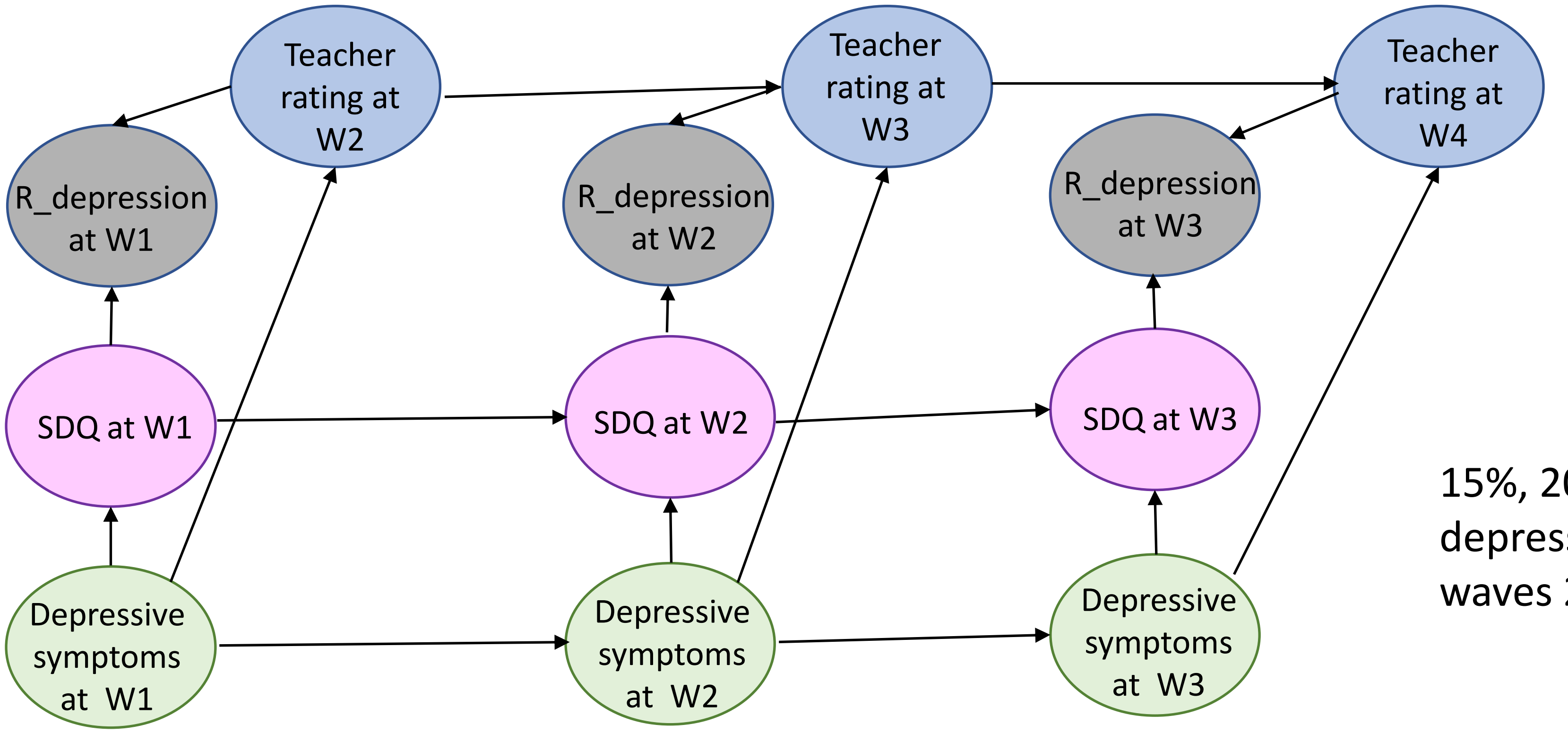
Combinations of cluster correlation at school (high = 0.15 and low = 0.01) and individual level (high = 0.5 and low = 0.2)

- High-high
- High-low
- Low-high
- Low-low

Simulation Study

Generation of missing data

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

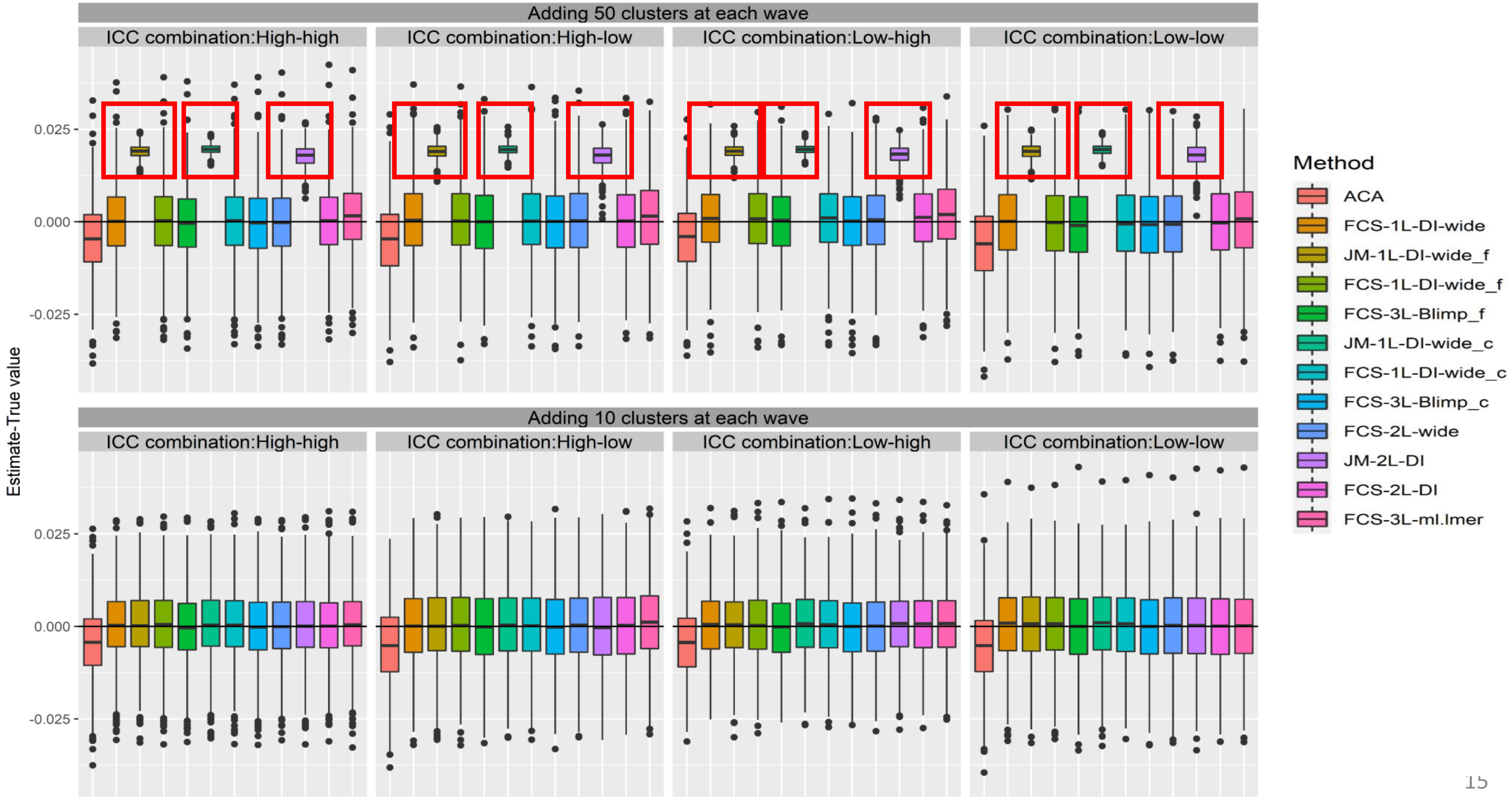


15%, 20% and 30% of the depressive symptom scores at waves 2,4, and 6 set to missing

Simulation results

Deviations from the true value- $\beta_1 = (-0.02)$

MAR-CATS scenario

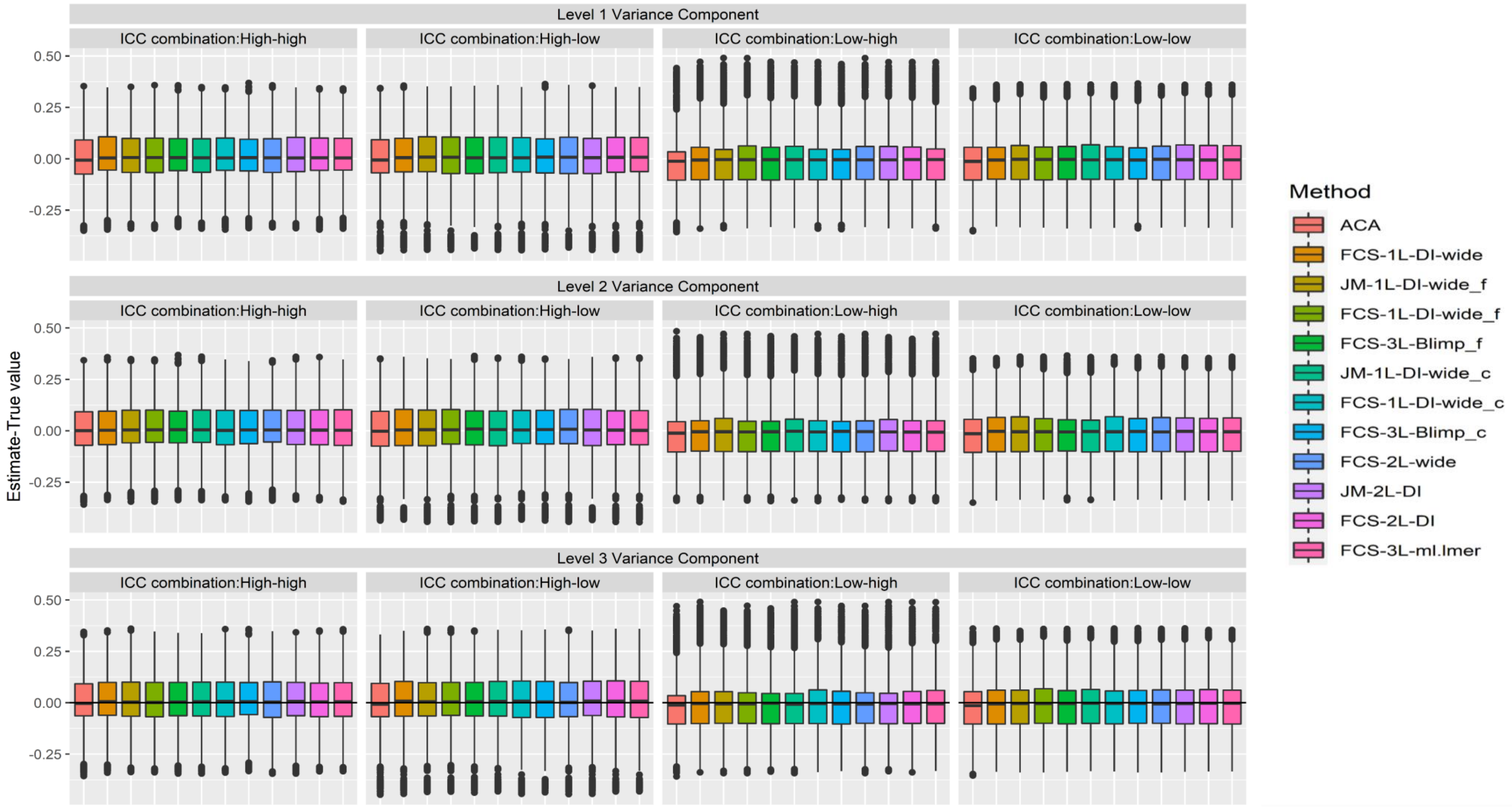


Simulation results

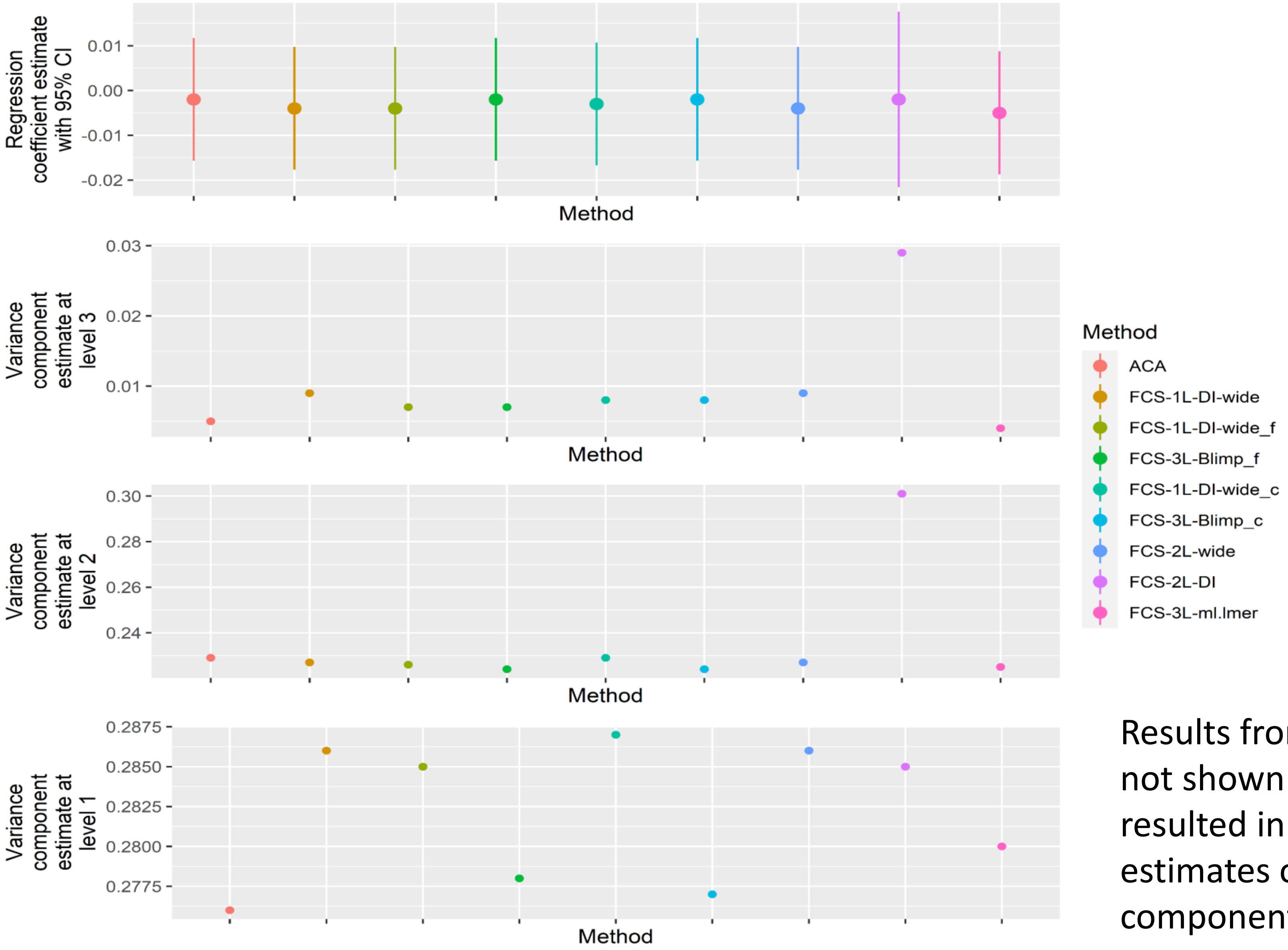
Deviations from the true value- variance components

MAR-CATS scenario

Adding 50 clusters at each wave



Case study illustration



Results from JM approaches not shown as they all resulted in implausible estimates of variance components

Conclusions

- The extensions of the single- 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 JM approaches extended with DI with caution as the large number of DIs in these approaches can lead to biased estimates of the regression coefficient and the variance components

Thank you

Find me at:

Virtual room OC5F

https://www.rwijesuriya.com/

@rush_099