Understanding the consequences of collinearity for multilevel models: The importance of disaggregation across levels

Haley E. Yaremych

Kristopher J. Preacher

Department of Psychology & Human Development, Vanderbilt University

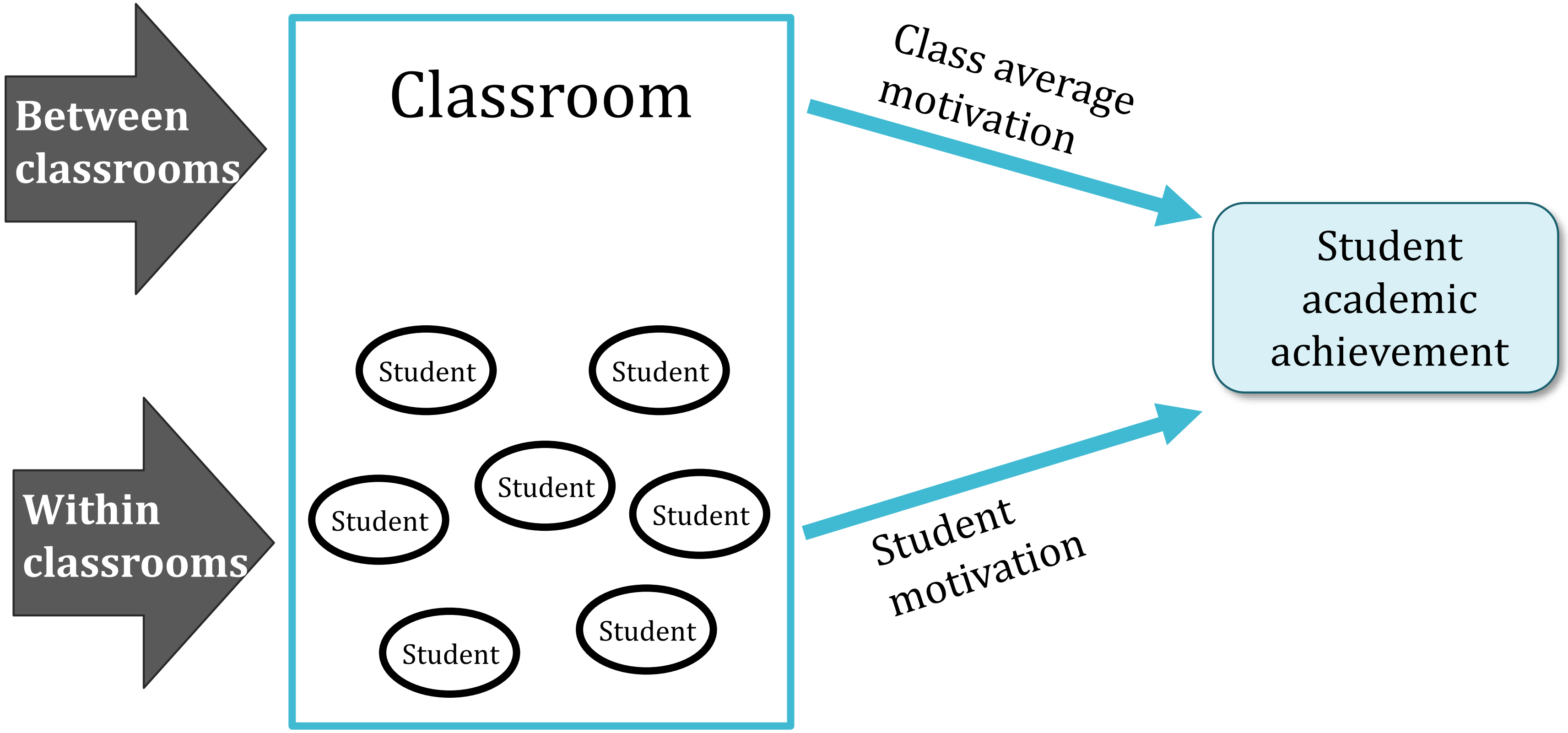
VANDERBILT UNIVERSITY

NSF

Outline

- **Background**
- Analytics
- Simulation
- Diagnostics
- Conclusions

Level-specific effects in multilevel data



The fully disaggregated model

The fully disaggregated model:

$$y_{ij} = \beta_{0j} + \beta_{1j}\left(x_{1ij} - \overline{x}_{1.j}\right) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{x}_{1.j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Reduced form:

$$\begin{aligned} y_{ij} = {} & \gamma_{00} + \gamma_{10}\left(x_{1ij} - \overline{x}_{1.j}\right) + \gamma_{01}\overline{x}_{1.j} \\ & + u_{0j} + u_{1j}\left(x_{1ij} - \overline{x}_{1.j}\right) + e_{ij} \end{aligned}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

The conflated model

The conflated model:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

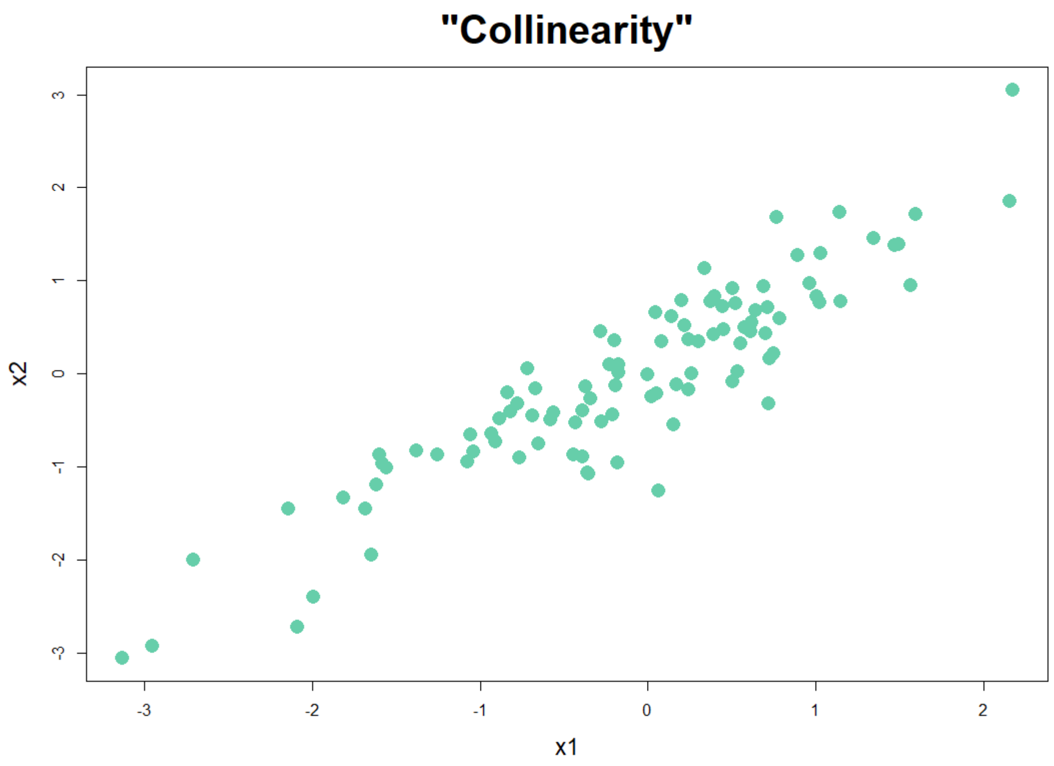
Reduced form:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

Collinearity



Consequences for single-level regression:

- Very large SEs (reduced power)
- Unstable point estimates
- Substantively impossible point estimates

Collinearity in multilevel data

- Consensus across (very few) studies:
  - Enlarged *SE*s of fixed effect estimates
  - Larger sample size helps
- Contradictory results across (very few) studies:
  - ICC_y
  - Relative bias in random effect (co)variance estimates
- Many unanswered questions, such as:
  - How do the consequences of collinearity change across different centering specifications (e.g., the conflated vs. the disaggregated model)?

Clark (2013); Hendrickx (2018); Shieh & Fouladi (2003); Stinnett (1993); Yu et al. (2015)

Study goals

- ***Analytics:*** Establish the consequences of collinearity for the conflated model
- ***Simulation:*** Clarify how the consequences of collinearity change across model specifications and data characteristics
- ***Diagnostics:*** Demonstrate how collinearity diagnostics are influenced by centering and disaggregation

Outline

- Background
- **Analytics**
- Simulation
- Diagnostics
- Conclusions

The generalized least squares (GLS) estimator

$$\hat{\beta}_{GLS} = \left\{ \sum_{j=1}^{J} \frac{X_{Bj}^T X_{Bj}}{1+(n_j-1)\rho} + \sum_{j=1}^{J} \frac{X_{Wj}^T X_{Wj}}{1-\rho} \right\}^{-1} \times \left\{ \sum_{j=1}^{J} \frac{X_{Bj}^T Y_j}{1+(n_j-1)\rho} + \sum_{j=1}^{J} \frac{X_{Wj}^T Y_j}{1-\rho} \right\}$$

$\hat{\beta}_{GLS}$ = a single conflated slope estimate

J = number of clusters

n_j = cluster size for cluster j

ρ = intraclass correlation of y_{ij}

X_{Bj} = a vector of cluster means for cluster j

X_{Wj} = a vector of cluster mean centered predictors for cluster j

Scott & Holt (1982)

GLS estimator: maximally general form

$$
\hat{\boldsymbol{\beta}}_{\mathbf{GLS}} = \left\{ \sum_{j=1}^{J} \left[\begin{array}{c:c} \left(1+\left(n_j-1\right)\rho\right)^{-1} n_j & \left(1+\left(n_j-1\right)\rho\right)^{-1} n_j \bar{\mathbf{x}}_j' \\ \hdashline \left(1+\left(n_j-1\right)\rho\right)^{-1} n_j \bar{\mathbf{x}}_j & \left(1+\left(n_j-1\right)\rho\right)^{-1} n_j \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j' + (1-\rho)^{-1}\left(\mathbf{X}_j'\mathbf{X}_j - n_j \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j'\right) \end{array} \right] \right\}^{-1}
$$

$$
\times \left\{ \sum_{j=1}^{J} \left(\left(1+(n_j-1)\rho\right)^{-1} \left[\mathbf{1}_{n_j} \;\vdots\; \mathbf{1}_{n_j} \bar{\mathbf{x}}_j' \right]^T Y_j + (1-\rho)^{-1} \left[\mathbf{0}_{n_j} \;\vdots\; \mathbf{X}_j - \mathbf{1}_{n_j} \bar{\mathbf{x}}_j' \right]^T Y_j \right) \right\}
$$

$\hat{\boldsymbol{\beta}}_{\mathbf{GLS}}$ = a vector of conflated slope estimates

J = number of clusters

n_j = cluster size for cluster j

ρ = intraclass correlation of y_{ij}

$\mathbf{1}_{\mathbf{n_j}}$ = $n_j \times 1$ column vector of 1's

$\mathbf{0}_{\mathbf{n_j}}$ = $n_j \times 1$ column vector of 0's

$\bar{\mathbf{x}}_{\mathbf{j}}$ = column vector of cluster means for cluster j

$\mathbf{X}_{\mathbf{j}}$ = original data matrix for cluster j

$\mathbf{Y}_{\mathbf{j}}$ = column vector of outcomes for cluster j

The GLS estimator is informed by predictor covariance

Muthén (1990):

$$\mathbf{S}_B = (J-1)^{-1}\sum_{j=1}^{J} n_j\left(\bar{x}_{.j}-\bar{x}_{..}\right)\left(\bar{x}_{.j}-\bar{x}_{..}\right)' \qquad \mathbf{S}_{PW} = (N-J)^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}_{.j}\right)\left(x_{ij}-\bar{x}_{.j}\right)'$$

Expressing components of the maximally general GLS estimator in terms of these quantities:

$$\sum_{j=1}^{J}\left(1+\left(n_j-1\right)\rho\right)^{-1} n_j\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j' = \left(1+\left(n-1\right)\rho\right)^{-1} n\begin{bmatrix} J & \sum_{j=1}^{J}\bar{\mathbf{z}}_j' \\ \sum_{j=1}^{J}\bar{\mathbf{z}}_j & \left(J-1\right)n^{-1}\color{red}{\mathbf{S}_B} \end{bmatrix}$$

$$\sum_{j=1}^{J}\left(1-\rho\right)^{-1}\left(\mathbf{X}_j'\mathbf{X}_j - n_j\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j'\right) = \left(1-\rho\right)^{-1}\begin{bmatrix} 0 & \sum_{j=1}^{J}\left(\mathbf{Z}_j - n\bar{\mathbf{z}}_j'\right) \\ \sum_{j=1}^{J}\left(\mathbf{Z}_j' - n\bar{\mathbf{z}}_j\right) & \left(N-J\right)\color{red}{\mathbf{S}_{PW}} \end{bmatrix}$$

Key takeaways

- Predictor covariance (a.k.a. predictor collinearity) systematically influences slope estimates in the conflated multilevel model
- Departure from single-level regression

Outline

- Background
- Analytics
- **Simulation**
- Diagnostics
- Conclusions

Simulation study design

Held constant:

- Three continuous level-1 predictors: $x_{1ij}, x_{2ij}, x_{3ij}$
- Continuous level-1 outcome: y_{ij}
- Within- and between-cluster effects
- Total $var(y_{ij})$
- Sample size at each level

Varied:

r_W: within-cluster $cor(x_{1ij}, x_{2ij})$	–0.9, –0.8, –0.7, –0.6, 0, 0.6, 0.7, 0.8, 0.9
r_B: between-cluster $cor(x_{1ij}, x_{2ij})$	–0.9, –0.8, –0.7, –0.6, 0, 0.6, 0.7, 0.8, 0.9
ICC_y	0.05, 0.3
$\mathrm{ICC}_{x_{1ij}}$, $\mathrm{ICC}_{x_{2ij}}$	0.05, 0.3

Simulation study design

- Conflated model
  - Level 1: $x_{1ij}, x_{2ij}, x_{3ij}$
- Fully disaggregated model
  - Level 1:$(x_{1ij} - \bar{x}_{1.j}), (x_{2ij} - \bar{x}_{2.j}), (x_{3ij} - \bar{x}_{3.j})$
  - Level 2: $\bar{x}_{1.j}, \bar{x}_{2.j}, \bar{x}_{3.j}$
- Partially disaggregated model
  - Level 1: $x_{1ij}, (x_{2ij} - \bar{x}_{2.j}), (x_{3ij} - \bar{x}_{3.j})$
  - Level 2: $x_{2.j}, x_{3.j}$
- Outcomes:
  - Fixed effect estimates
  - Relative bias in the fixed effect estimate *SE*s
  - Relative bias in the random effect (co)variance estimates

Results: conflated model

- Both r_W and r_B influenced fixed effect estimates

- Fixed effect estimates were most strongly influenced when…
  - ICC_y was small
  - $\mathrm{ICC}_{x_{1ij}}$ and $\mathrm{ICC}_{x_{2ij}}$ were large

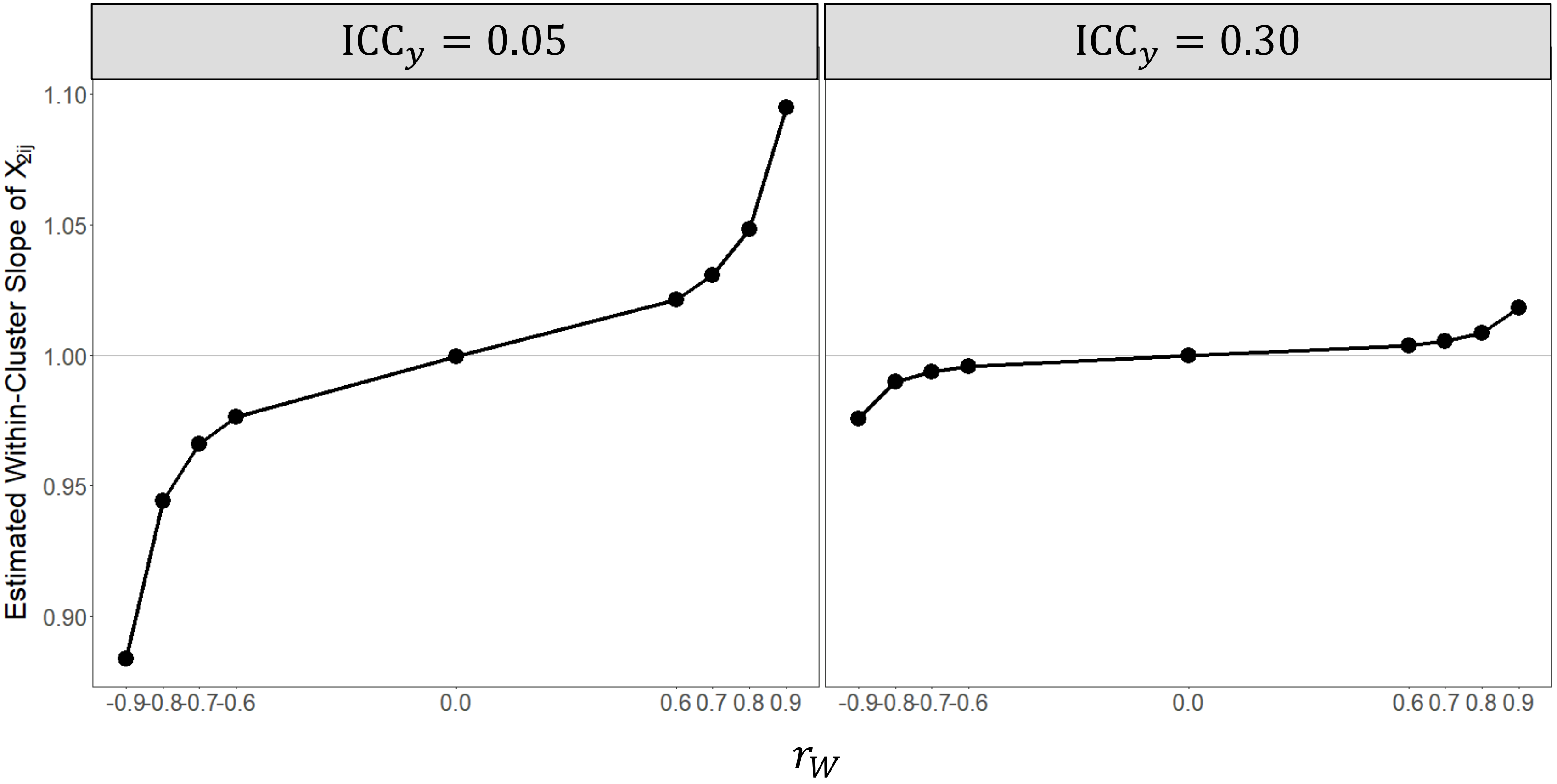
- Did not examine *SE*s or random effect (co)variance estimates due to problematic nature of fixed effect estimates

Results: partially disaggregated model

- Conflated slope of x_{1ij}:
  - Not affected
- Within-cluster slope of x_{2ij}:
  - $r_W \times \mathrm{ICC}_y$
  - $r_W \times$ predictor ICCs
- Between-cluster slope of x_{2ij}:
  - $r_B \times$ predictor ICCs

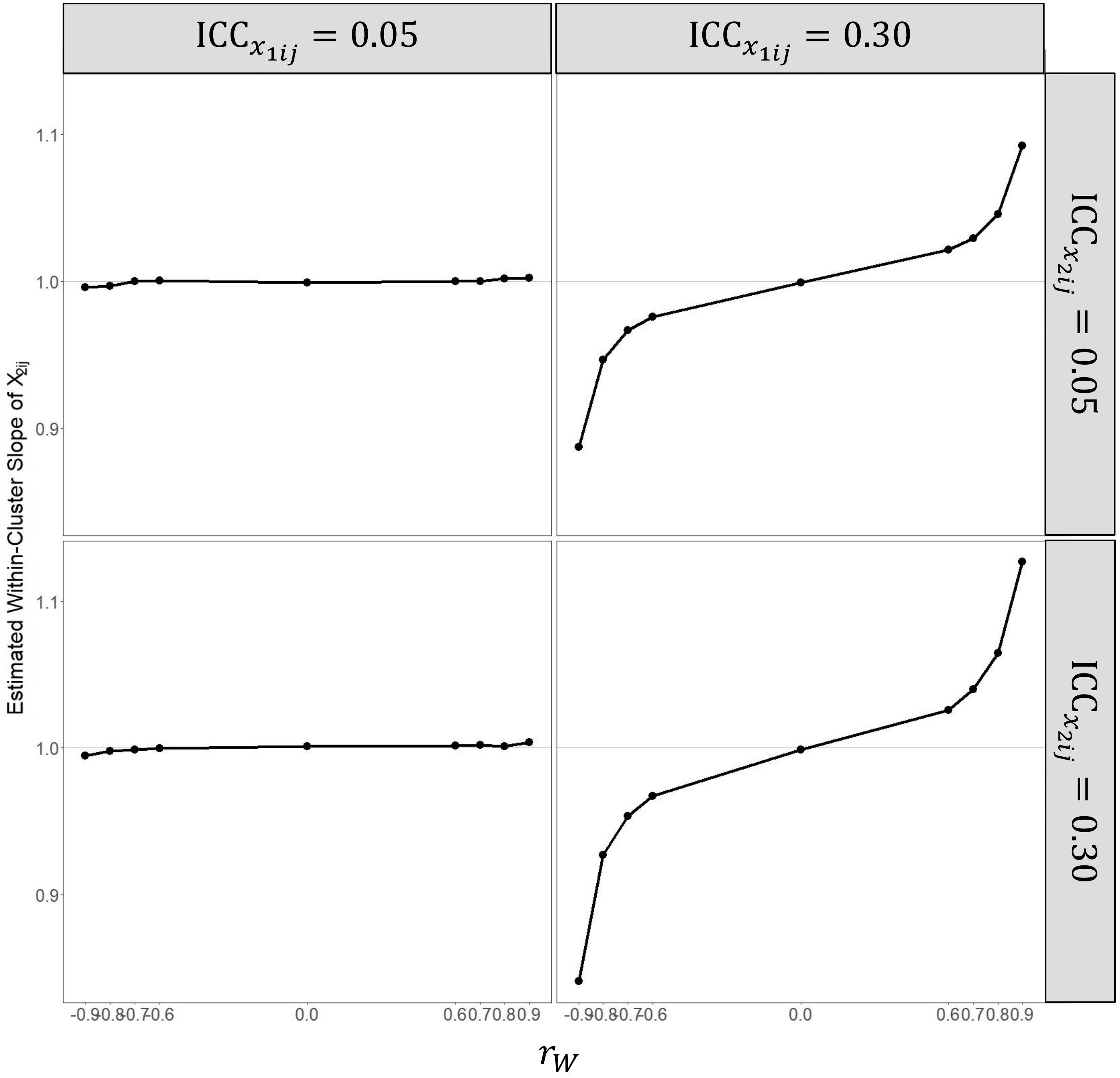
Results: partially disaggregated model

Estimated within-cluster slope of x_{2ij}



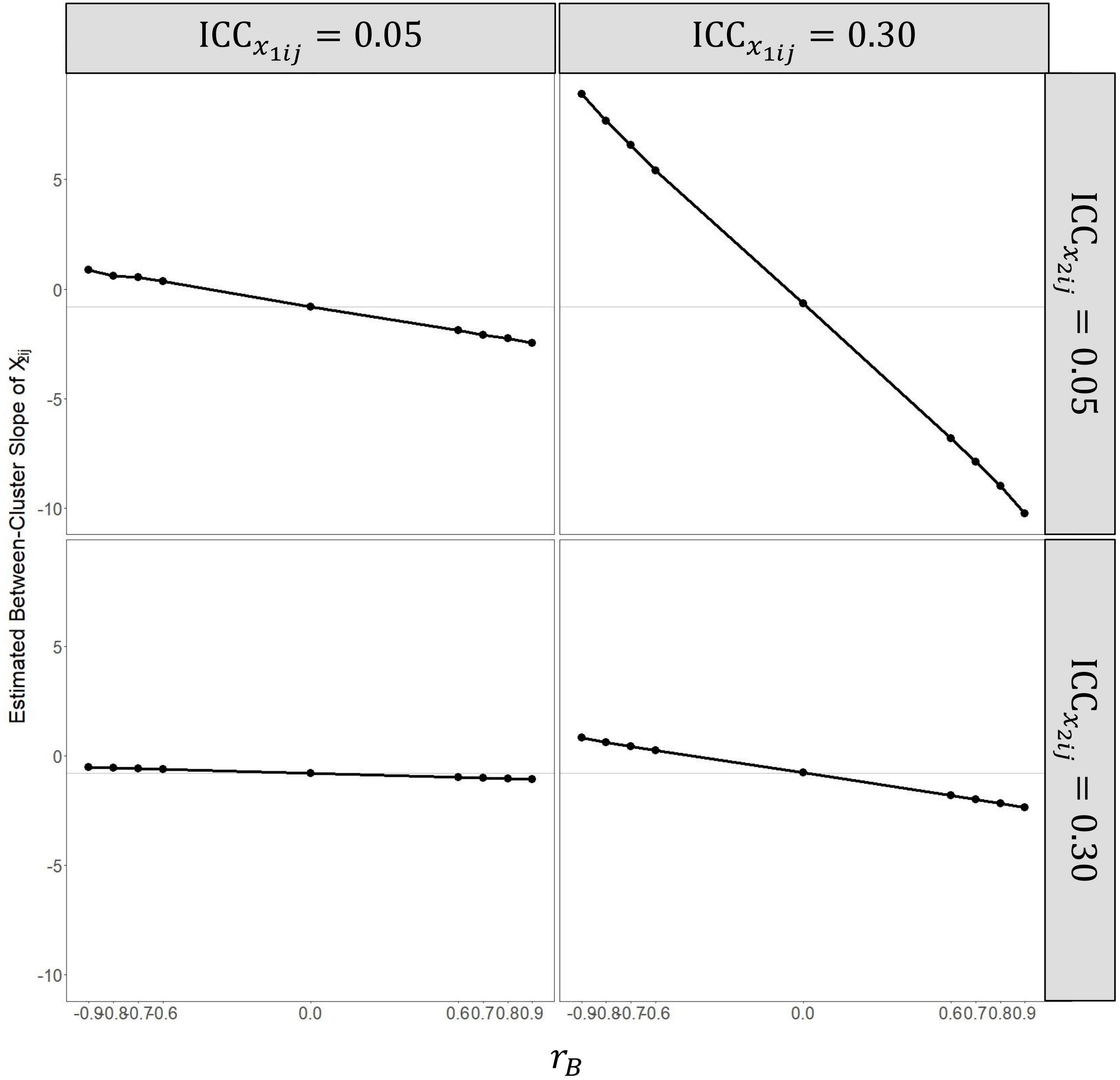
Results: partially disaggregated model

Estimated within-cluster slope of x_{2ij}



Results: partially disaggregated model

Estimated between-cluster slope of x_{2ij}



Takeaways from the partially disaggregated model

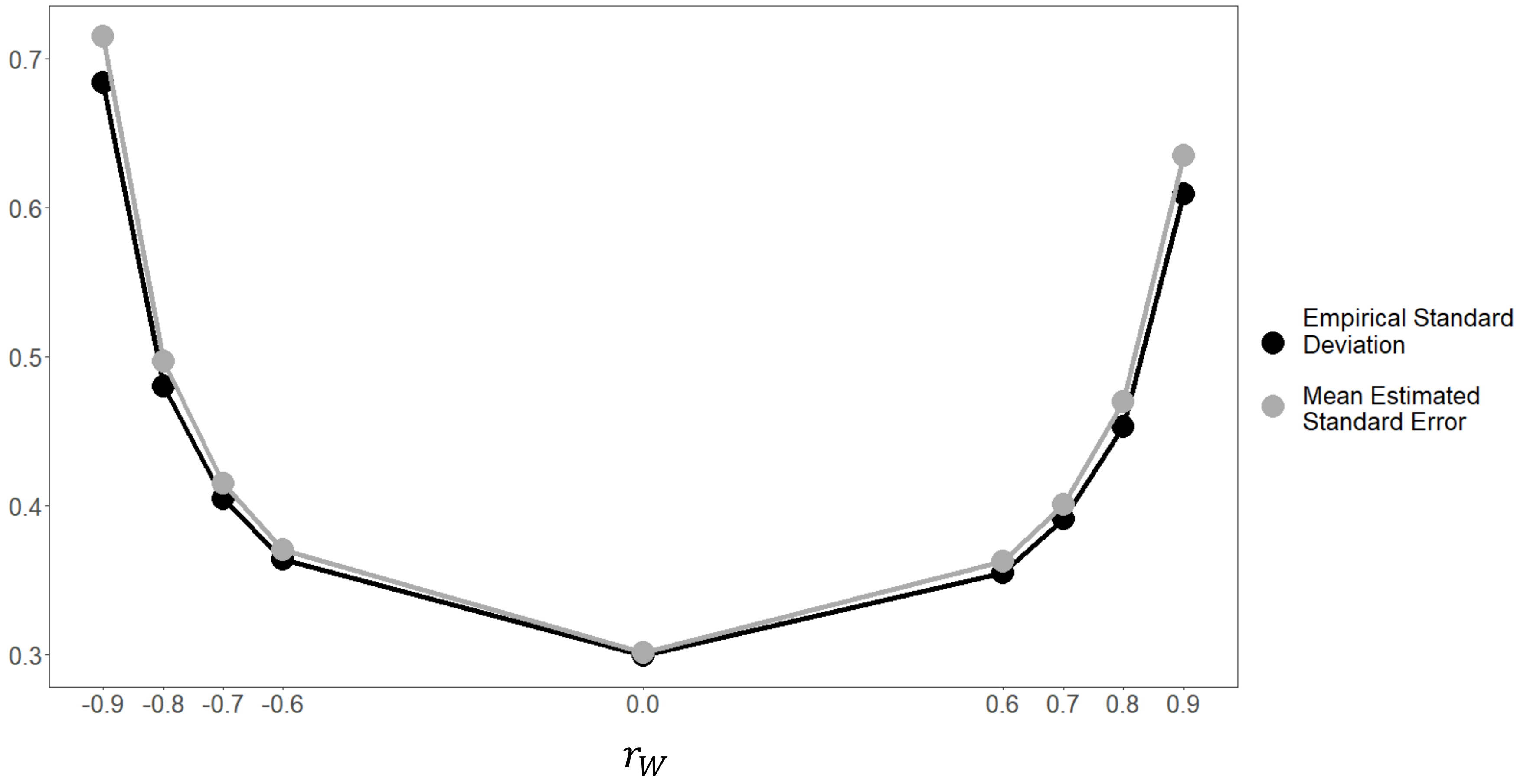
- When predictors are collinear and some are left uncentered whereas other, likely the most substantively important, predictors are disaggregated….
- That disaggregation will *not* always yield unbiased estimates!

Results: fully disaggregated model

- Fixed effect estimates
  - Unbiased in all conditions
- Relative bias in the *SE*s of fixed effect estimates
  - Within-cluster estimates: main effect of r_W
- Relative bias in the random effect (co)variance estimates
  - r_W interacted with all other design factors

Results: fully disaggregated model

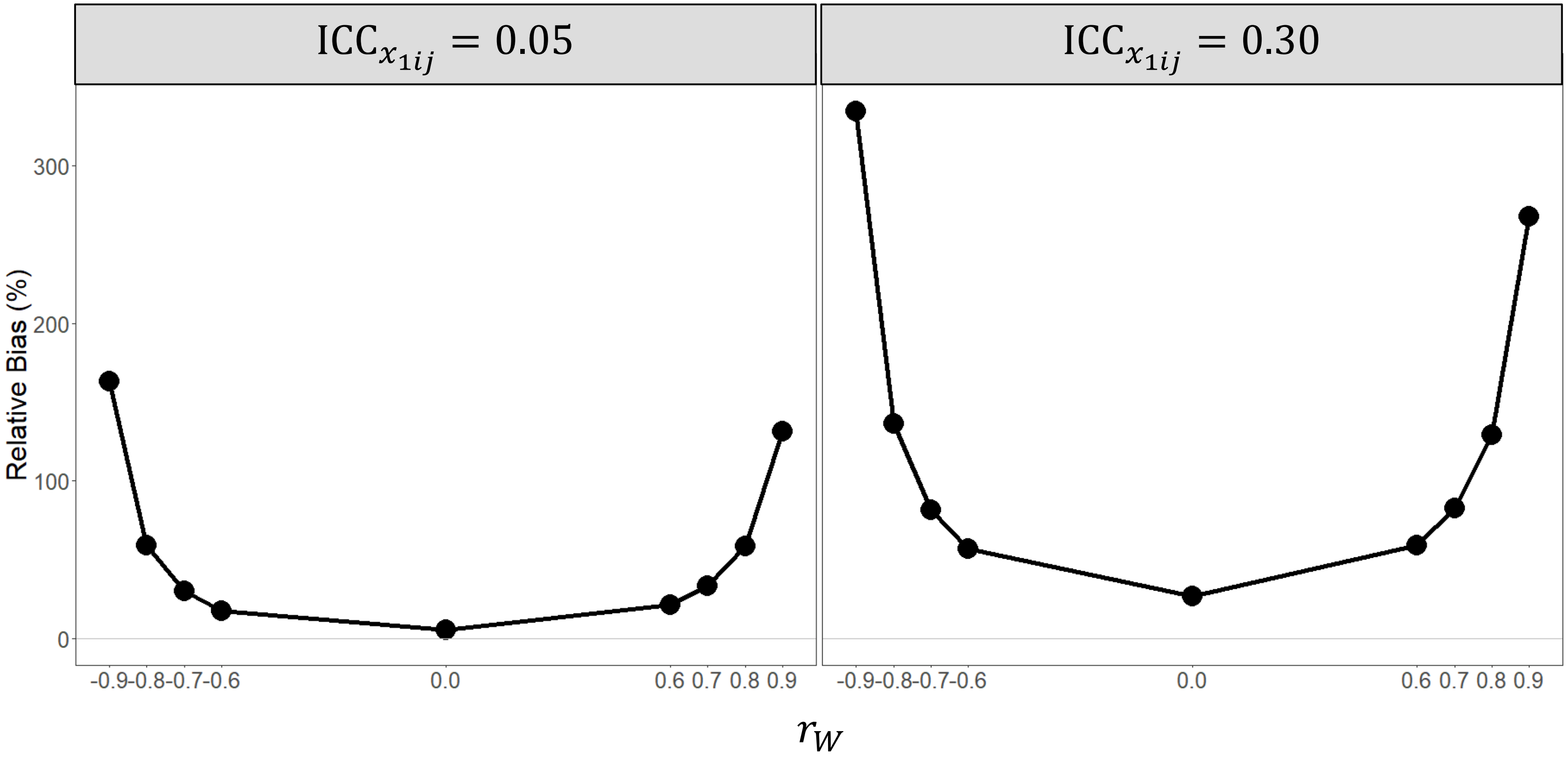
Standard error of the within-cluster effect of x_{1ij}



Results: fully disaggregated model

Random slope variance estimates

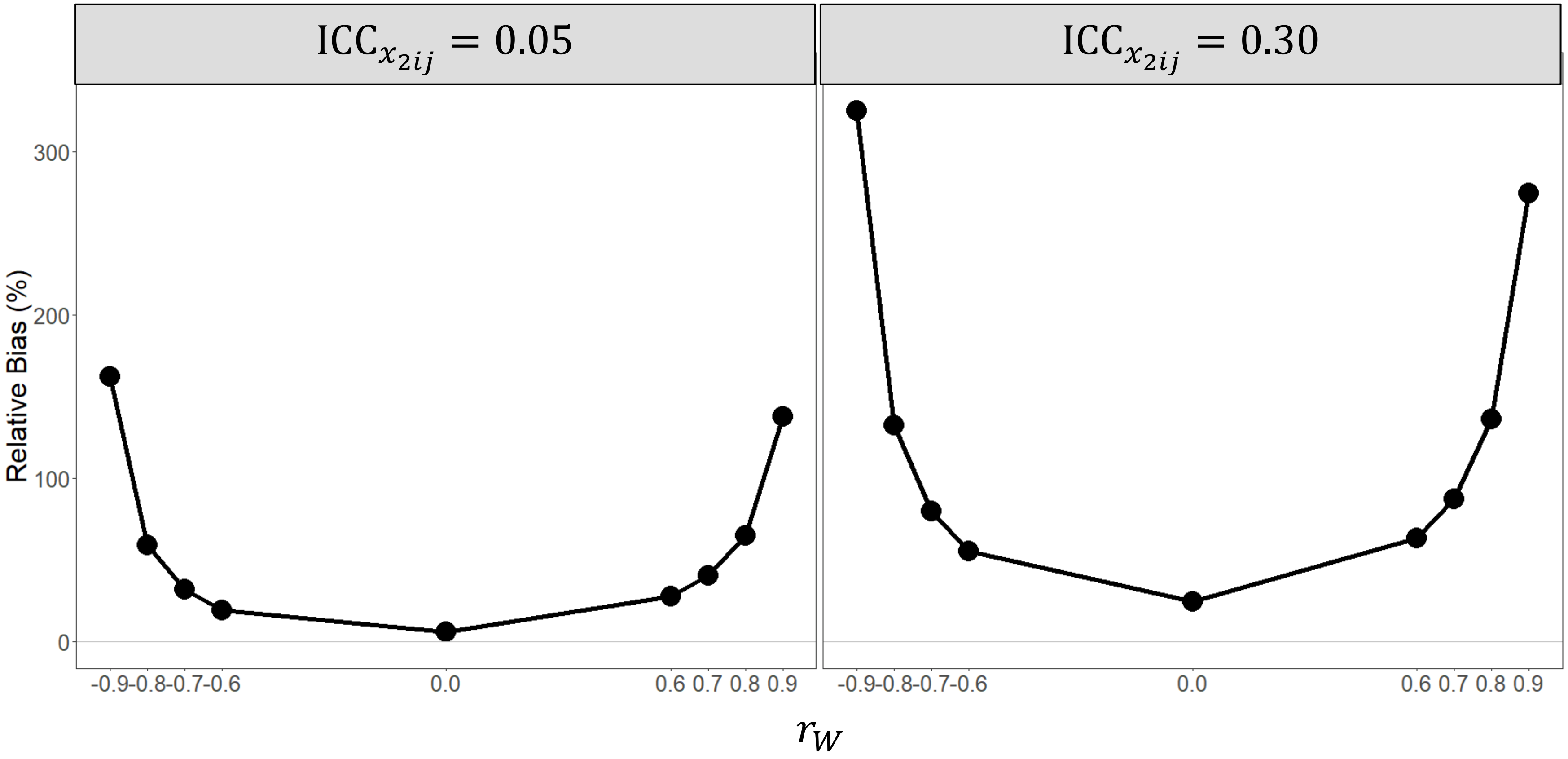
Relative bias in τ_{11}



Results: fully disaggregated model

Random slope variance estimates

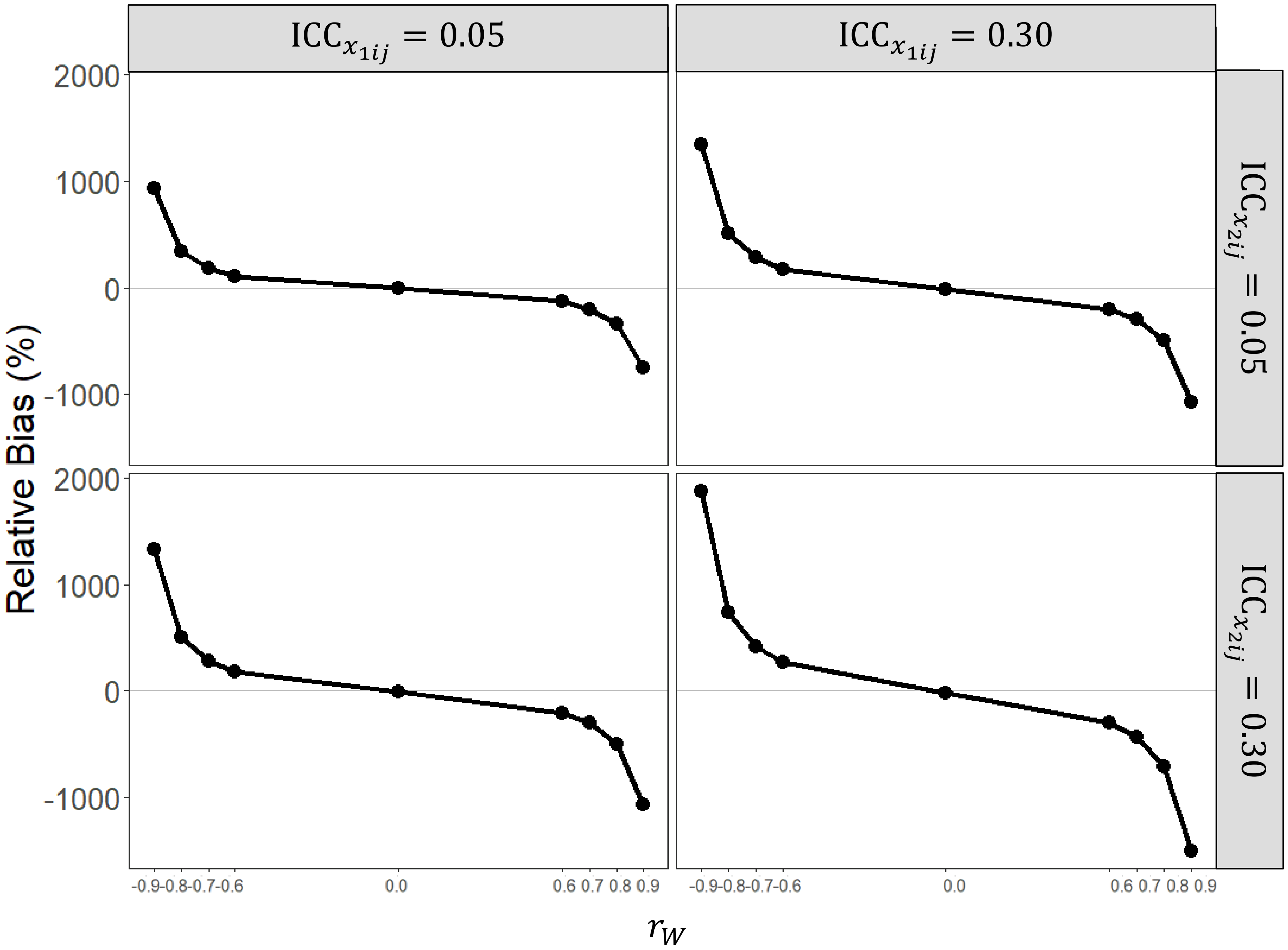
Relative bias in τ_{22}



Relative bias in τ_{21}

Results: fully disaggregated model

Random slope covariance estimates



Takeaways from the fully disaggregated model

- Results mimicked single-level regression
  - Point estimates unaffected
  - Standard errors enlarged

- Standard errors:
  - True increase in variability
  - ***AND*** upward bias in estimated *SE*s

- Random effect (co)variance estimates:
  - Influenced by r_W
  - Often extremely biased

Outline

- Background
- Analytics
- Simulation
- **Diagnostics**
- Conclusions

Collinearity diagnostics

- Variance Inflation Factor (VIF)
  - Comes from the formula for the variance of a slope estimate (single-level regression)
  - Interpretation: multiplicative factor by which $var(\beta_i)$ is increased due to collinearity in the data set
  - $var(\beta_i) = \frac{var_y}{var_{x_i}}\left(\frac{1-R_y^2}{n-k-1}\right)\left(\frac{1}{1-R_{x_i}^2}\right)$

- Condition number (κ)
  - Do an eigen-decomposition of $\boldsymbol{X'X}$ and obtain eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

  - Interpretation: “sensitivity” of regression results to small changes in the data set

Simulated data

Held constant:

- Three continuous level-1 predictors: $x_{1ij}, x_{2ij}, x_{3ij}$
- x_{3ij} was minimally correlated with x_{1ij} and x_{2ij}

Varied:

- Within- and between-cluster correlation of x_{1ij} and x_{2ij} $(\boldsymbol{r_W}, \boldsymbol{r_B})$
- $\text{ICC}_{x_{1ij}}$ and $\text{ICC}_{x_{2ij}}$
- These factors incidentally vary the "total" correlation of x_{1ij} and x_{2ij} $(\boldsymbol{r_T})$

Diagnostics computed on uncentered & level-disaggregated predictor sets:

- VIF
- Condition number (κ)

$r_W = 0, r_B = 0, \mathrm{ICC}_{x_{1ij}} = \mathrm{ICC}_{x_{2ij}} = 0.25$

Data set #1

Predictor set	**Correlation matrix**	**VIFs**	**Condition numbers (κs)**
Disaggregated	$\begin{array}{lllllll} & x_{1i} & x_{2i} & x_{3i} & x_{1.j} & x_{2.j} & x_{3.j} \\ x_{1i} & 1 & & & & & \\ x_{2i} & \mathbf{-0.023} & 1 & & & & \\ x_{3i} & 0.041 & 0.021 & 1 & & & \\ x_{1.j} & 0 & 0 & 0 & 1 & & \\ x_{2.j} & 0 & 0 & 0 & \mathbf{-0.095} & 1 & \\ x_{3.j} & 0 & 0 & 0 & -0.016 & 0.112 & 1 \end{array}$	x_{1i}: 1.002 x_{2i}: 1.001 x_{3i}: 1.002 $x_{1.j}$: 1.009 $x_{2.j}$: 1.022 $x_{3.j}$: 1.013	Level 1: 1.051 Level 2: 1.158
Uncentered	$\begin{array}{llll} & x_{1ij} & x_{2ij} & x_{3ij} \\ x_{1ij} & 1 & & \\ x_{2ij} & \mathbf{-0.031} & 1 & \\ x_{3ij} & 0.036 & 0.030 & 1 \end{array}$	x_{1ij}: 1.002 x_{2ij}: 1.002 x_{2ij}: 1.002	1.053

$r_W = 0.7, r_B = -0.9, \mathrm{ICC}_{x_{1ij}} = \mathrm{ICC}_{x_{2ij}} = 0.25$

Data set #2

Predictor set	Correlation matrix	VIFs	Condition numbers (κs)
Disaggregated	$\begin{array}{lllllll} & x_{1i} & x_{2i} & x_{3i} & x_{1.j} & x_{2.j} & x_{3.j} \\ x_{1i} & 1 & & & & & \\ x_{2i} & \mathbf{0.720} & 1 & & & & \\ x_{3i} & 0.083 & 0.085 & 1 & & & \\ x_{1.j} & 0 & 0 & 0 & 1 & & \\ x_{2.j} & 0 & 0 & 0 & \mathbf{-0.905} & 1 & \\ x_{3.j} & 0 & 0 & 0 & -0.036 & -0.086 & 1 \end{array}$	x_{1i}: 2.080 x_{2i}: 2.081 x_{3i}: 1.008 $x_{1.j}$: 5.959 $x_{2.j}$: 5.995 $x_{3.j}$: 1.085	Level 1: 2.494 Level 2: 4.687
Uncentered	$\begin{array}{llll} & x_{1ij} & x_{2ij} & x_{3ij} \\ x_{1ij} & 1 & & \\ x_{2ij} & \mathbf{0.509} & 1 & \\ x_{3ij} & 0.066 & 0.060 & 1 \end{array}$	x_{1ij}: 1.353 x_{2ij}: 1.352 x_{2ij}: 1.005	1.763

r_W = 0.25, r_B = 0.95, $\text{ICC}_{x_{1ij}}$= 0.8, $\text{ICC}_{x_{2ij}}$= 0.01

Data set #3

Predictor set	**Correlation matrix**	**VIFs**	**Condition numbers (κs)**
Disaggregated	$\begin{array}{lllllll} & x_{1i} & x_{2i} & x_{3i} & x_{1.j} & x_{2.j} & x_{3.j} \\ x_{1i} & 1 & & & & & \\ x_{2i} & \mathbf{0.254} & 1 & & & & \\ x_{3i} & 0.074 & 0.048 & 1 & & & \\ x_{1.j} & 0 & 0 & 0 & 1 & & \\ x_{2.j} & 0 & 0 & 0 & \mathbf{0.955} & 1 & \\ x_{3.j} & 0 & 0 & 0 & 0.309 & 0.316 & 1 \end{array}$	x_{1i}: 1.073 x_{2i}: 1.070 x_{3i}: 1.006 $x_{1.j}$: 11.328 $x_{2.j}$: 11.380 $x_{3.j}$: 1.111	Level 1: 1.311 Level 2: 6.864
Uncentered	$\begin{array}{llll} & x_{1ij} & x_{2ij} & x_{3ij} \\ x_{1ij} & 1 & & \\ x_{2ij} & \mathbf{0.073} & 1 & \\ x_{3ij} & 0.120 & 0.046 & 1 \end{array}$	x_{1ij}: 1.019 x_{2ij}: 1.007 x_{2ij}: 1.016	1.153

Key takeaways

- Collinearity diagnostics applied to uncentered predictors are misleading and arbitrary
- Level-specific collinearity influences bias and precision in all models investigated here (*even the fully conflated model!*)
- In all cases, level-specific collinearity must be diagnosed in order to understand how estimation has been impacted

Outline

- Background
- Analytics
- Simulation
- Diagnostics
- **Conclusions**

Conclusions

- To ensure that point estimates will not be biased due to collinearity, ***disaggregate all predictors!***

- Depending on data conditions, expect that collinearity may introduce bias into *SE*s and/or random effect estimates
  - Potential avenues for mitigation…
  - Larger ICC_y
  - Smaller predictor ICCs

- Limitations
  - Many design factors were held constant

Unanswered questions

- Diagnosing collinearity in multilevel data
  - Accepted cutoffs?
  - Performance?

- Optimal strategies for remedying collinearity problems in multilevel data
  - Removing the predictor(s) with strongest collinearity?
  - Multilevel PCA?
  - Multilevel factor analysis?

References

- Clark, P. C. (2013). The effects of multicollinearity in multilevel models (Doctoral dissertation). Wright State University.
- Hendrickx, J. (2018). *Collinearity in mixed models*. Paper presented at PHUSE EU Connect Conference, Frankfurt, Germany.
- Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, *77*(380), 848-854.
- Shieh, Y. Y., & Fouladi, R. T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement*, *63*(6), 951-985.
- Stinnett, S. S. (1994). *Collinearity in mixed models*. (Doctoral dissertation). The University of North Carolina at Chapel Hill.
- Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, 53, 118-136.

Thank you!

haley.e.yaremych@vanderbilt.edu

VANDERBILT UNIVERSITY

NSF