# Relationships between centering choices and collinearity problems in multilevel models

Haley E. Yaremych & Kristopher J. Preacher

Department of Psychology & Human Development, Vanderbilt University

VANDERBILT UNIVERSITY

## Background

### *Centering & Conflation in Multilevel Models*

- Uncentered (UN) level-1 predictors yield slope estimates that are *conflated*, uninterpretable mixes of within- and between-cluster effects.
- In contrast, inclusion of the cluster mean as a level-2 predictor alongside the cluster-mean-centered (CWC) level-1 predictor is often advocated. This approach effectively separates the unique within- and between-cluster effects of the predictor, yielding an *unconflated* model.
- Raudenbush & Bryk (2002)[1] derived an equation to algebraically predict the conflated estimate that would arise from a single UN predictor:

$$\hat{\gamma}^*_{10} = \frac{W_1\hat{\beta}_b + W_2\hat{\beta}_w}{W_1 + W_2} \quad ; \quad W_1 = \left[\mathrm{var}\left(\hat{\beta}_b\right)\right]^{-1} \quad ; \quad W_2 = \left[\mathrm{var}\left(\hat{\beta}_w\right)\right]^{-1}$$

- The conflated estimate is a precision-weighted average of its within- and between-cluster effects. It is unknown whether this equation holds for multiple predictors, which typically covary at both levels.

### *Multicollinearity in Multilevel Models*

- Sparse prior work suggests multicollinearity causes similar problems in single- and multilevel settings (unstable point estimates, large SEs of fixed effects)[2,3,4]
- None of this work has addressed collinearity problems as they relate to centering choices. It is unknown whether different centering choices may exaggerate or mitigate the harmful effects of collinearity.

## Aims & Hypotheses

***Aim:*** to determine whether different centering choices for level-1 predictors yield models that differ in susceptibility to the harmful effects of multicollinearity in MLM.

***Hypothesis:*** In general, conflated estimates will be more susceptible to the harmful effects of collinearity than unconflated (i.e., level-specific) estimates. Specifically, conflated point estimates will change as the strength and direction (positive/negative) of predictor correlation changes at both levels.

## Analytics

- Our goal was to analytically show whether and how covariance among level-1 predictors would impact conflated slope estimates. Because maximum likelihood (ML) estimates are algebraically intractable, we turned to the the generalized least squares ($\beta_{GLS}$) estimator[5]. $\beta_{GLS}$ is asymptotically equivalent to ML.
- We derived the maximally general form of the $\beta_{GLS}$ estimator, which allows for any number of predictors, and any number of clusters of potentially varying sizes:

$$\hat{\beta}_{GLS} = \left\{\sum_{j=1}^{J}\left[\begin{array}{c|c}\left(1+(n_j-1)\rho\right)^{-1}n_j & \left(1+(n_j-1)\rho\right)^{-1}n_j\bar{\mathbf{x}}'_j \\ \hline \left(1+(n_j-1)\rho\right)^{-1}n_j\bar{\mathbf{x}}_j & \left(1+(n_j-1)\rho\right)^{-1}n_j\bar{\mathbf{x}}_j\bar{\mathbf{x}}'_j+(1-\rho)^{-1}\left(\mathbf{X}'_j\mathbf{X}_j-n_j\bar{\mathbf{x}}_j\bar{\mathbf{x}}'_j\right)\end{array}\right]\right\}^{-1} \times \left\{\sum_{j=1}^{J}\left(\left(1+(n_j-1)\rho\right)^{-1}\left[\mathbf{1}_{n_j} \mid \mathbf{1}_{n_j}\bar{\mathbf{x}}'_j\right]^T Y_j + (1-\rho)^{-1}\left[\mathbf{0}_{n_j} \mid \mathbf{X}_j - \mathbf{1}_{n_j}\bar{\mathbf{x}}'_j\right]^T Y_j\right)\right\}$$

$\rho$ = intraclass correlation of $y_{ij}$

$\bar{\mathbf{x}}_j$ = vector of predictor means in cluster $j$

$\mathbf{X}_j$ = matrix of predictor values in cluster $j$

$n_j$ = cluster size

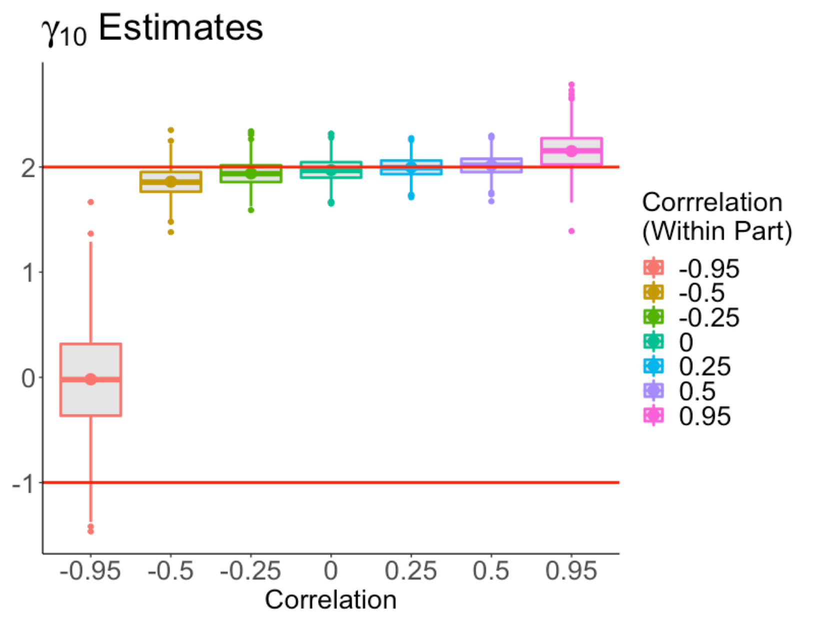
## Simulation Study

- Multilevel data sets (100 clusters, each of size 30, $ICC_Y$ = .3; 1000 data sets per condition) with two level-1 predictors and a level-1 outcome were simulated. Correlation at level 1, $cor(x_{1ij}-x_{1.j}, x_{2ij}-x_{2.j})$, and level 2, $cor(x_{1.j}, x_{2.j})$, was varied while correlation at the other level was held at zero.
- We then fit three models: (1) *fully conflated,* where both $x_{1ij}$ and $x_{2ij}$ were uncentered; (2) *partially conflated*, where $x_{1ij}$ was split into level-specific parts and $x_{2ij}$ was uncentered; (3) *unconflated*, where both $x_{1ij}$ and $x_{2ij}$ were split into level-specific parts. Point estimates and their SEs were recorded.
- Point estimates associated with $x_{1ij}$ are shown below. Its true within-cluster effect was 2 (upper red line), and its true between-cluster effect was –1 (lower red line).

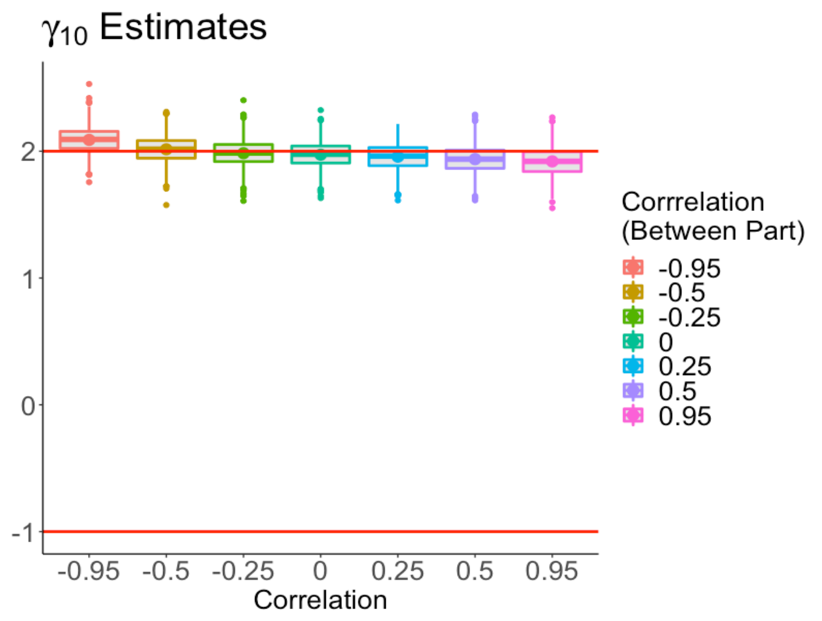
### *Fully Conflated Model*

$$y_{ij} = \gamma^*_{00} + \gamma^*_{10}x_{1ij} + \gamma^*_{20}x_{2ij} + u^*_{0j} + e^*_{ij}$$

*Varying level-1 correlation*



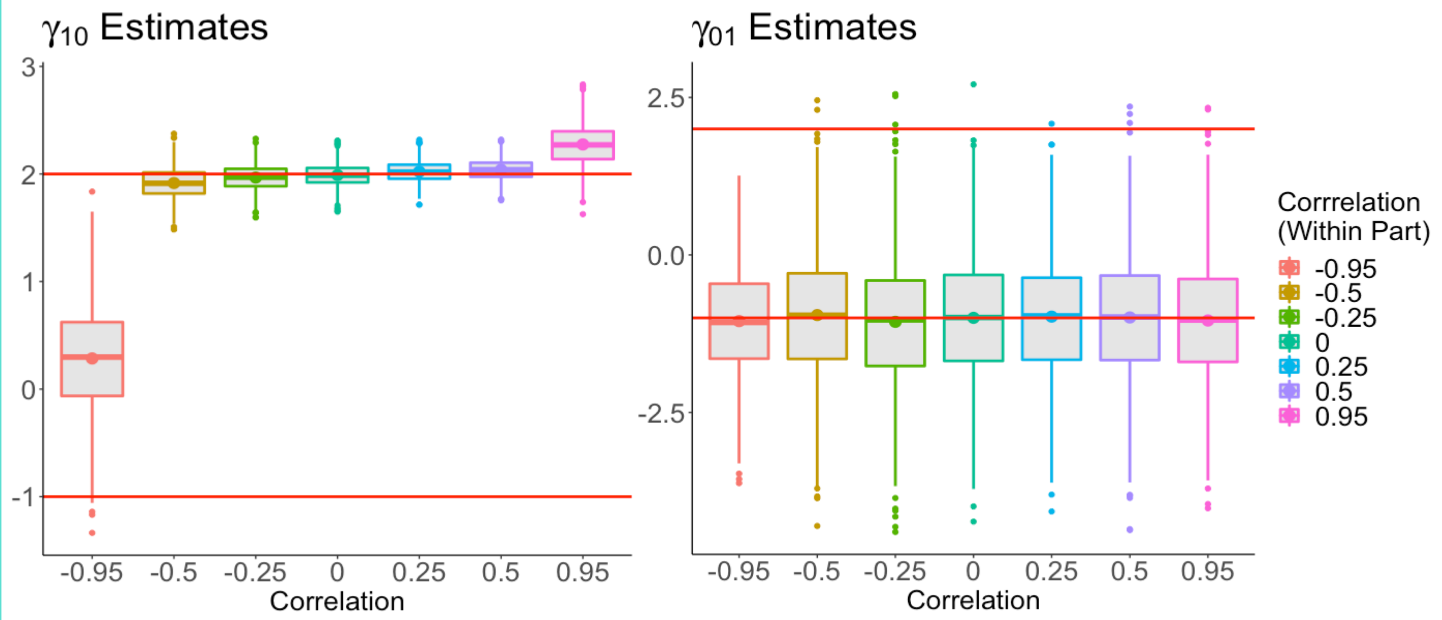

*Varying level-2 correlation*



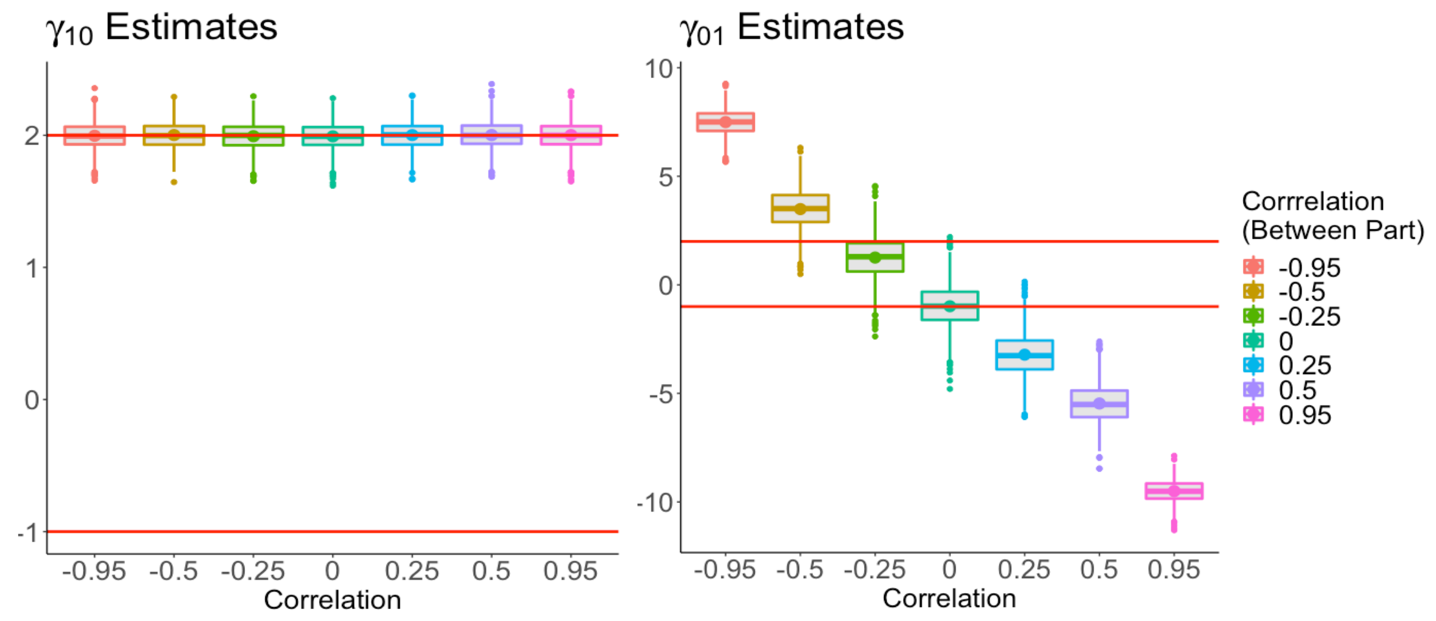

### *Partially Conflated Model*

$$y_{ij} = \gamma_{00} + \gamma_{10}\left(x_{1ij}-x_{1.j}\right) + \gamma_{01}x_{1.j} + \gamma^*_{20}x_{2ij} + u_{0j} + e_{ij}$$

*Varying level-1 correlation*



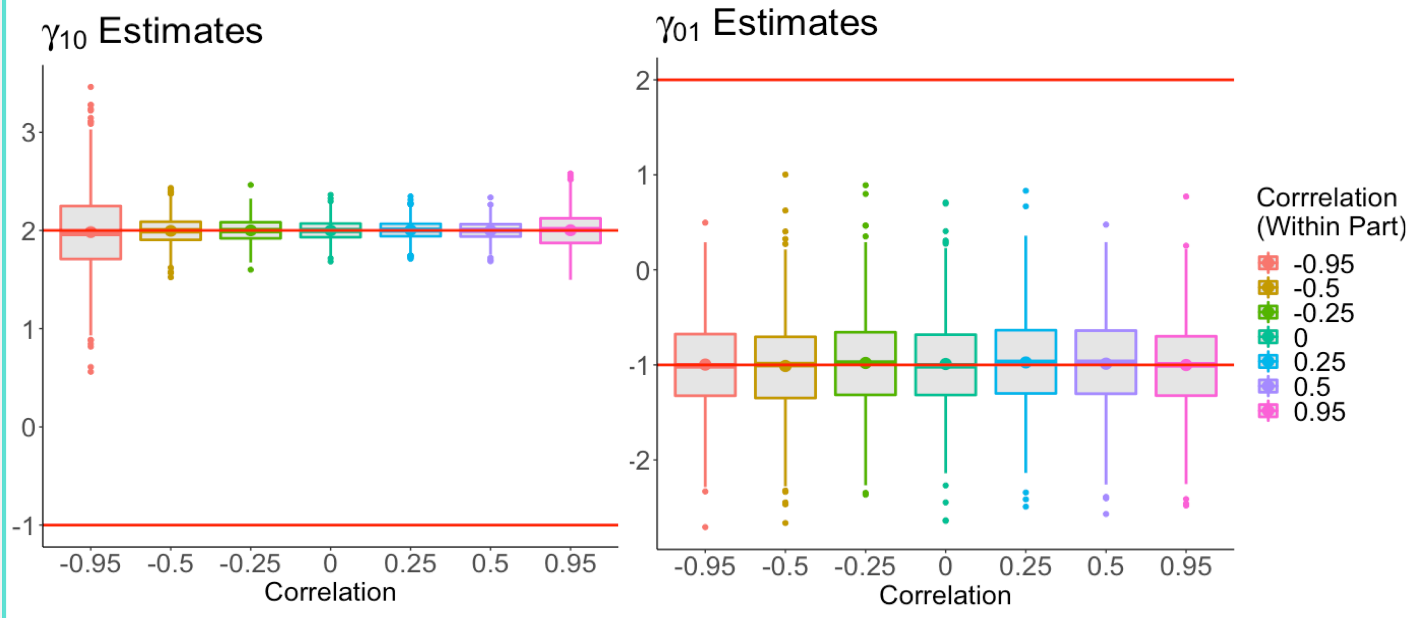

*Varying level-2 correlation*



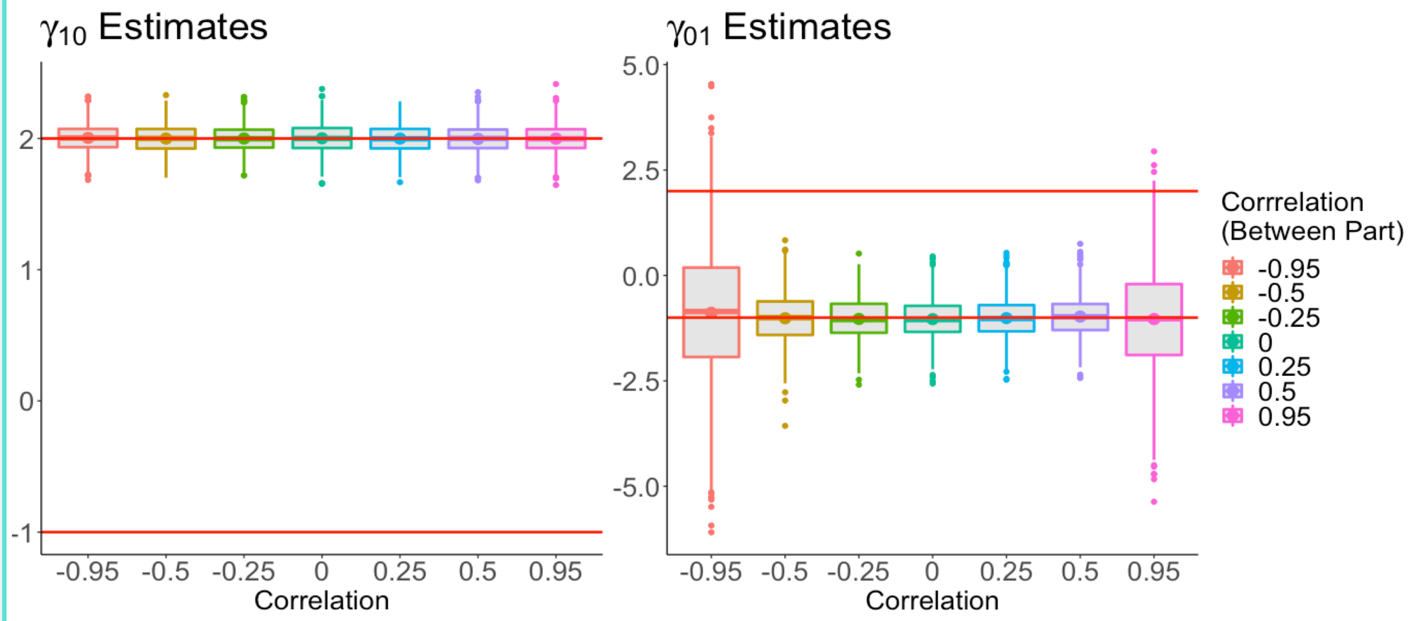

### *Unconflated Model*

$$y_{ij} = \gamma_{00} + \gamma_{10}\left(x_{1ij}-x_{1.j}\right) + \gamma_{01}x_{1.j} + \gamma_{20}\left(x_{2ij}-x_{2.j}\right) + \gamma_{02}x_{2.j} + u_{0j} + e_{ij}$$

*Varying level-1 correlation*




*Varying level-2 correlation*




## Conclusions

- Our derivation of the $\beta_{GLS}$ estimator shows that each conflated slope estimate varies as a function of within- and between-cluster covariance among predictors.
- In contrast, unconflated point estimates are robust to inaccurate point estimates as a result of collinearity.
- Interestingly, in the partially conflated model, level-specific point estimates still varied as a function of predictor covariance.
  - ➢ This suggests that inclusion of *any* uncentered predictors may result in bias that propagates throughout the model, the severity of which is a function of collinearity strength.
- The unconflated model still suffered from large SEs when collinearity at the relevant level was extremely strong, but did not suffer from biased point estimates under any condition.

## Next Steps

- Expand the simulation study to examine how other data characteristics (e.g., cluster size, number of clusters, $ICC_X$), interact with collinearity to exaggerate or mitigate its harmful effects.
- Record the degree of mismatch between observed conflated estimates and those predicted by the Raudenbush & Bryk (2002) equation.
- Evaluate the utility of diagnostic measures (e.g., kappa coefficient, multilevel VIF) for identifying problematic levels of collinearity in multilevel models.[2,3]

## References

1. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

2. Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, *53*, 118-136.

3. Clark, P. C., Jr. (2013). *The effects of multicollinearity in multilevel models* (Doctoral dissertation, Wright State University).

4. 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.

5. 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.