What if Sample Size is the Confound in Multilevel Models?

Michael S. Truong, Cathy Xijuan Zhang and David B. Flora

York University

July 15/2025

Contact: emstruonger@gmail.com

Today's Plan

Conclusion

The traditional presentation of the role of sample size in statistics is inadequate: a naive reader of the replication crisis may believe that all their problems will be solved with a large enough sample size and enough high quality measures.

Such a belief is wrong. Data alone cannot solve your problems. (Pearl & Mackenzie, 2018)

The good news is that this presentation shows that when the role of sample size is correctly specified, statistical problems more or less vanish. Therefore, investigators should carefully balance the goal of maximizing sample size against the unknown effects of sample size in their problem.

Contact: emstruonger@gmail.com

Context: Sample Size Effects Are Underappreciated and Important

What Should We Consider When We Plan our Sample Sizes?

Modern Issues:

1. Replication Crisis: Sample Size $\uparrow$ = Science $\uparrow$ (Collaboration, 2015; Lakens, 2022; Pargent et al., 2024)
   - ~~Money, Time and Resources~~
2. Psychology of Individuals v.s. Psychology of Average Individual (Molenaar, 2004)
   - Intensive Longitudinal Data
   - Deep Phenotyping
   - Increase psychometric reliability (Hedge et al., 2018)

However, what if sample size is a confound?

- What if the process of measurement itself is a causal factor?
- Parallel: Longitudinal Measurement Invariance (McNeish et al., 2021; Telzer et al., 2018; Vogelsmeier et al., 2024)

How Does Sample Size Relate to Multilevel Modeling? And What Does This Have to do with Sample Size Effects?

MLMs weighs $\widehat{CM}_{0j}$ between $\overline{CM}_{0j}$ and γ_{00}, based on amount of evidence (N_j):

$$\widehat{CM}_{0j} = \frac{\frac{N_j}{\sigma^2_{\epsilon}}\overline{CM}_{0j} + \frac{1}{\sigma^2_{cm}}\gamma_{00}}{\frac{N_j}{\sigma^2_{\epsilon}} + \frac{1}{\sigma^2_{cm}}} \quad (1)$$

Let's say that increasing classroom sizes causes class mean grades to increase, then what?

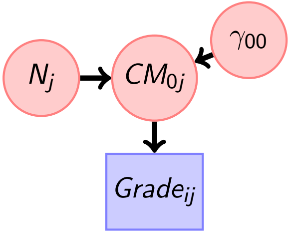
- **Usual:** Judgments (estimates) change with the evidence
- **Unusual:** Judgments (estimates) change with the ***amount*** of evidence, independently of weighing
- Will MLM explode???

(Gelman & Hill, 2007, Eq.18.12; McElreath, 2020)

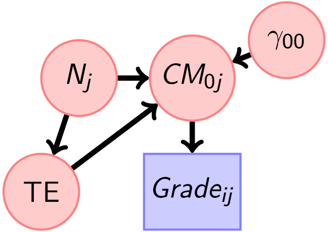
Let's investigate this issue using a simulation study on the effects of classroom size

Simulation Study on Sample Size Effects in Multilevel Models

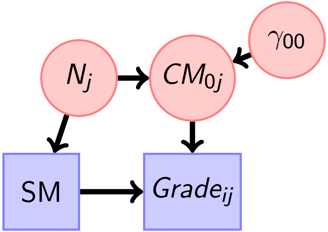
Method: 3 Data Generating Processes



(a) DGP1: Classroom Size on Class Mean Grade



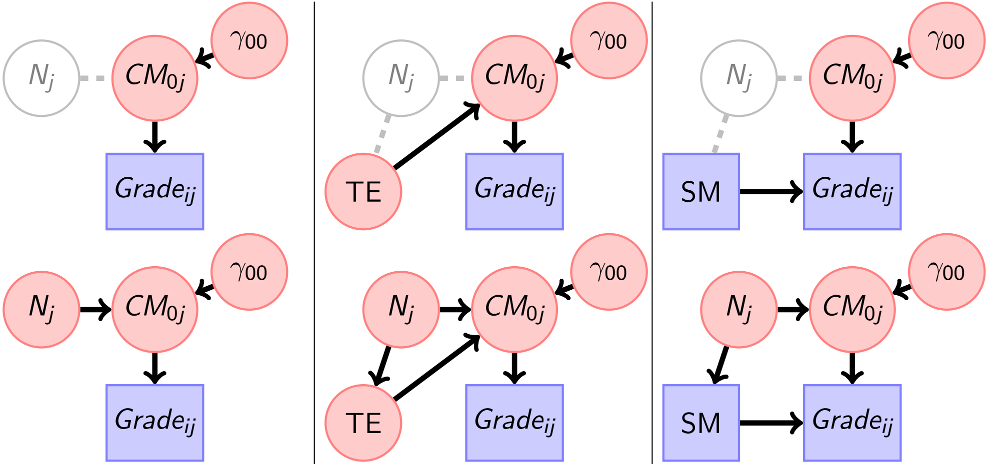
(b) DGP2: Classroom Size on Class Mean Grade and Teacher Experience



(c) DGP3: Classroom Size on Class Mean Grade and Student Motivation

- Truncate CM_{0j} and $Grade_{ij} \in [0, 100]$

Method: Estimate 2 Models per Data Generating Process



(a) Estimated DGP1 Models (b) Estimated DGP2 Models (c) Estimated DGP3 Models

Objective: When we vary the effects of classroom size and vary whether we control classroom size effects, how will our parameter estimates be affected?

Results

Outline of Results

- Bias and RMSE
  - Controlling classroom size $\rightarrow$ unbiased
  - Bias-Variance tradeoff with # of clusters
  - Lower level estimates unaffected
- Standard Errors
  - Biased when # of classrooms $\downarrow$
    - Non-zero classroom size effect
    - Controlling classroom size
  - Lower level estimates unaffected
- Anomalous Results
  - Condition-specific
  - Likely due to heteroscedasticity & truncation

Bias and RMSE: General

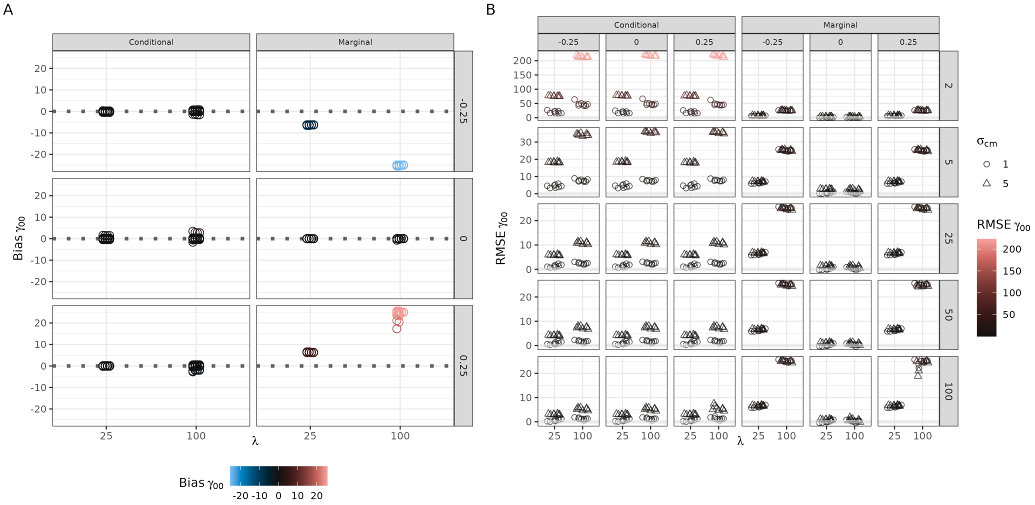


Figure 3: DGP1 γ_{00}. RMSE columns by γ_{cs}, rows by J

Bias and RMSE: Lower-Level Estimates

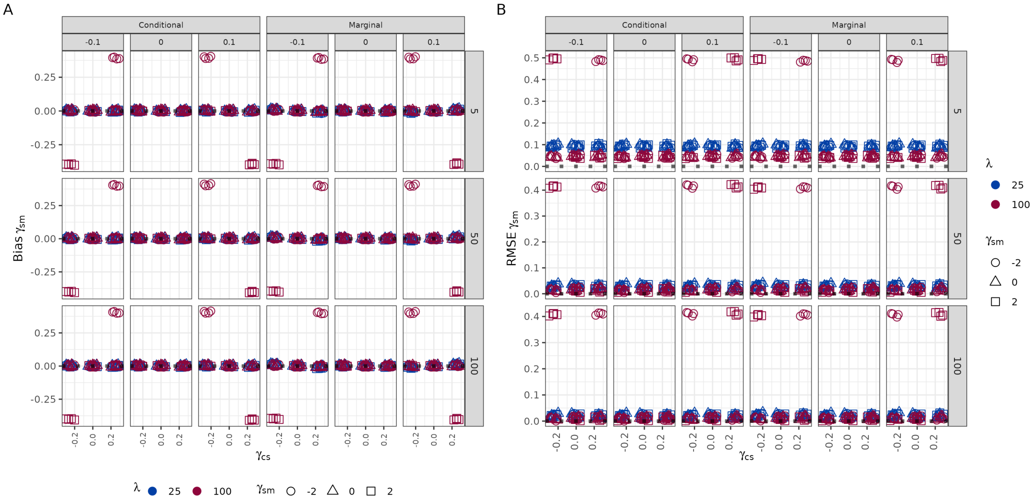


Figure 4: DGP3 γ_{sm}. Columns by $\gamma_{cs\ on\ sm}$, rows by J

Standard Errors: General

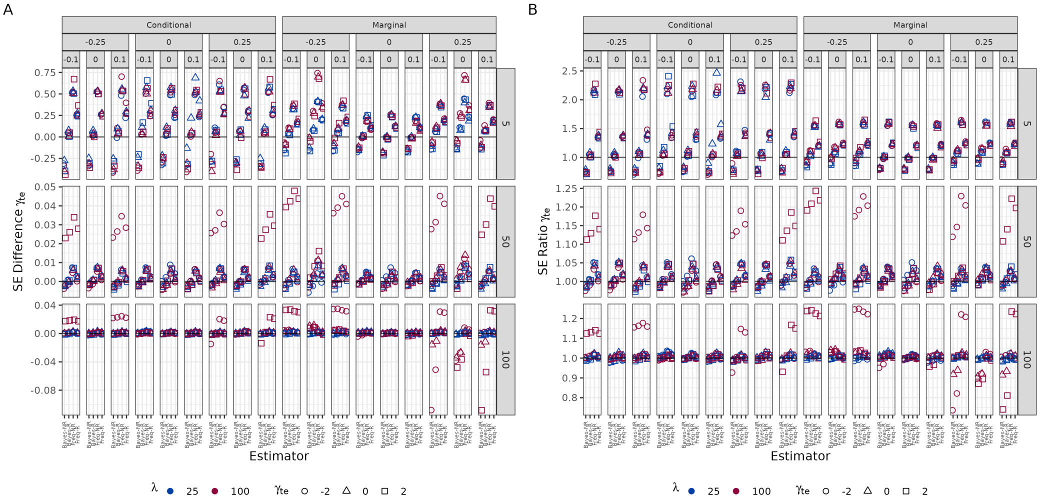


Figure 5: DGP2 Standard Error of $\widehat{\gamma}_{te}$. Columns by γ_{cs} and $\gamma_{cs\ on\ te}$, rows by J

Standard Errors: Lower-Level Estimates

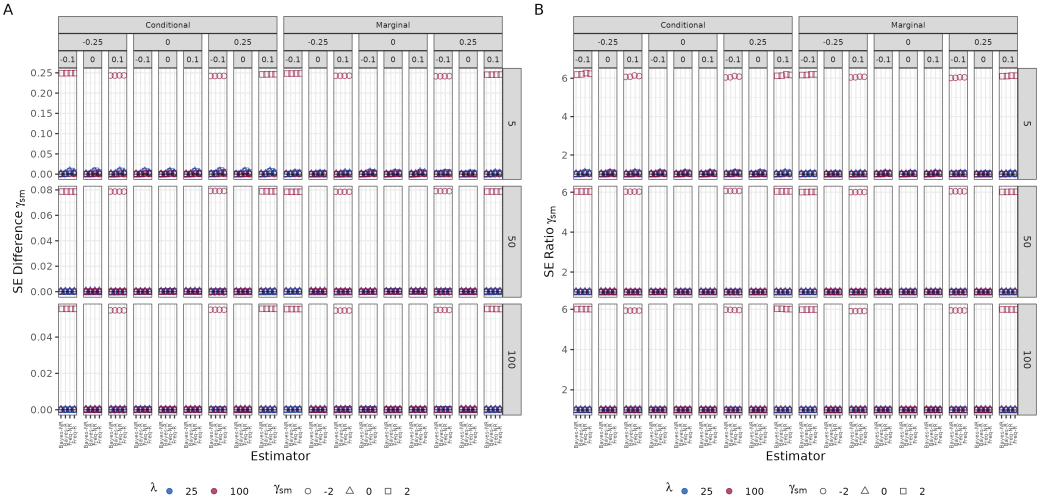


Figure 6: DGP3 Standard Error of $\widehat{\gamma}_{sm}$. Columns by γ_{cs} and $\gamma_{cs\ on\ sm}$, rows by J

What do these simulations tell us?

What Should We Remember During Data Analysis? How Should We Analyze Our Data?

What to remember:

- With small # of clusters:
  - Bias-Variance tradeoff (Hastie et al., 2009; Raudenbush & Schwartz, 2020)
  - Inaccurate standard errors no matter what
- Lower-level estimates are generally unaffected due to random intercept (Cinelli et al., 2024)

What to do:

- **General Solution:** Sensitivity analyses and draw causal DAGs
- **Recommendation:** Simulation-Based Calibration and tailored sample-size (Gelman et al., 2020; Pargent et al., 2024; Talts et al., 2020)

How Should We Collect Our Data and Interpret Our Findings?

Data Collection:

- Problem of known vs unknown missing data
- Nonlinear cluster size effects → sample range of cluster sizes
- Field-specific guidelines: probability of sample size (cluster size) effect?[1]
- Unusually easy: heteroscedasticity and range-restriction in measurements

Interpretation of Findings:

- Nonlinear (moderation) effects of sample size (cluster size) effects may make design-based control **misleading**
- Consider sample size effects on structural and measurement invariance

[1] Special Attention: Psycho-physical and neuropsychological studies

How Can We Improve On This Research?

- Accuracy of random slope estimates? Intercept-slope correlation?
- Group-mean centering?
- Meta-Analysis? Meta-Regression?
- Cross-classified models?
- Instrumental variables as a solution? (Ehrenberg et al., 2001)
- Measurement error in sample size (cluster size)? (Ehrenberg et al., 2001)
- Robust standard errors?

Conclusion

The traditional presentation of the role of sample size in statistics is inadequate: a naive reader of the replication crisis may believe that all their problems will be solved with a large enough sample size and enough high quality measures.

Such a belief is wrong. Data alone cannot solve your problems. (Pearl & Mackenzie, 2018)

The good news is that this presentation shows that when the role of sample size is correctly specified, statistical problems more or less vanish. Therefore, investigators should carefully balance the goal of maximizing sample size against the unknown effects of sample size in their problem.

Contact: emstruonger@gmail.com

References I

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science [Publisher: American Association for the Advancement of Science Section: Research Article]. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement [Publisher: SAGE Publications Inc]. *Psychological Science in the Public Interest*, *2*(1), 1–30. https://doi.org/10.1111/1529-1006.003

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* [OCLC: ocm67375137]. Cambridge University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. https://doi.org/10.48550/arXiv.2011.01808

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Lakens, D. (2022). Sample size justification (D. v. Ravenzwaaij, Ed.). *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Taylor; Francis, CRC Press.

References II

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data [Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2021.1915788]. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever [Publisher: Routledge _eprint: https://doi.org/10.1207/s15366359mea0204_1]. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Pargent, F., Koch, T. K., Kleine, A.-K., Lermer, E., & Gaube, S. (2024). A tutorial on tailored simulation-based sample-size planning for experimental designs with generalized linear mixed models [Publisher: SAGE Publications Inc]. *Advances in Methods and Practices in Psychological Science*, *7*(4), 25152459241287132. https://doi.org/10.1177/25152459241287132

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Raudenbush, S. W., & Schwartz, D. (2020). Randomized experiments in education, with implications for multilevel causal inference [Publisher: Annual Reviews]. *Annual Review of Statistics and Its Application*, *7*, 177–208. https://doi.org/10.1146/annurev-statistics-031219-041205

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2020, October 21). Validating bayesian inference algorithms with simulation-based calibration. https://doi.org/10.48550/arXiv.1804.06788

Telzer, E. H., McCormick, E. M., Peters, S., Cosme, D., Pfeifer, J. H., & van Duijvenvoorde, A. C. K. (2018). Methodological considerations for developmental longitudinal fMRI research. *Developmental Cognitive Neuroscience*, *33*, 149–160. https://doi.org/10.1016/j.dcn.2018.02.004

References III

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research*, *33*(8), 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

DGP1 Mathematics

$$Grade_{ij} \sim \mathcal{N}(CM_{0j}, \sigma^2_\epsilon), \text{for } i = 1, \dots CS_j, \text{ and } j = 1, \dots J, \tag{2}$$

$$CM_{0j} \sim \mathcal{N}(\gamma_{00} + \gamma_{cs} \cdot CS_j, \sigma^2_{cm}), \text{for } j = 1, \dots J, \tag{3}$$

$$CS_j \sim \min\{Pois(\lambda), 1\}, \text{for } j = 1, \dots J, \tag{4}$$

$$grade_{ij} \sim \mathcal{N}\left(CM_{0j}, \sigma^2_\epsilon\right) \tag{5}$$

$$CM_{0j} \sim \mathcal{N}\left(\gamma_{00}, \sigma^2_{cm}\right) \tag{6}$$

$$grade_{ij} \sim \mathcal{N}\left(CM_{0j}, \sigma^2_\epsilon\right) \tag{7}$$

$$CM_{0j} \sim \mathcal{N}\left(\gamma_{00} + \gamma_{cs}\,\mathrm{CS}_j, \sigma^2_{cm}\right). \tag{8}$$

DGP1 Simulation Conditions I

Parameter	Description	Value
σ_ϵ	Error standard deviation.	$\{1, 5\}$
Number of Clusters (J)	Number of J clusters for the MCS replication	$\{2, 5, 25, 50, 100\}$
λ	Specifies the Poisson distribution to take one draw of to determine a given cluster's size.	$\{25, 100\}$
γ_{cs}	γ_{cs} point change in mean cluster grade per student.	$\{-.25, 0, .25\}$
σ_{cm}	Between cluster standard deviation	$\{1, 5\}$
γ_{00}	Population mean grade	$\{50\}$

DGP2 Mathematics

$$\begin{aligned}
Grade_{ij} &\sim \mathcal{N}(CM_{0j}, \sigma^2_{\epsilon}), \text{for } i = 1, \ldots CS_j, \text{ and } j = 1, \ldots J, && (9)\\
CM_{0j} &\sim \mathcal{N}(\gamma_{00} + \gamma_{cs} \cdot CS_j + \gamma_{te} TE_j, \sigma^2_{cm}), \text{for } j = 1, \ldots J, && (10)\\
TE_j &\sim \mathcal{N}(0 + \gamma_{\text{cs on te}} CS_j, \sigma^2_{te}), \text{for } j = 1, \ldots J, && (11)\\
CS_j &\sim \min\{Pois(\lambda), 1\}, \text{for } j = 1, \ldots J, && (12)
\end{aligned}$$

$$\begin{aligned}
grade_{ij} &\sim N\left(CM_{0j}, \sigma^2_{\epsilon}\right) && (13)\\
CM_{0j} &\sim N\left(\gamma_{00} + \gamma_{te} TE_j, \sigma^2_{cm}\right) && (14)
\end{aligned}$$

$$\begin{aligned}
grade_{ij} &\sim N\left(CM_{0j}, \sigma^2_{\epsilon}\right) && (15)\\
CM_{0j} &\sim N\left(\gamma_{00} + \gamma_{te} TE_j + \gamma_{cs} CS_j, \sigma^2_{cm}\right). && (16)
\end{aligned}$$

DGP2 Simulation Conditions I

Parameter	Description	Value
Number of Clusters (J)	Number of J clusters for the MCS replication	$\{5, 50, 100\}$
λ	Specifies the Poisson distribution to take one draw of to determine a given cluster's size.	$\{25, 100\}$
γ_{cs}	γ_{cs} point change in mean cluster grade per student.	$\{-.25, 0, .25\}$
γ_{te}	γ_{te} point change in mean cluster grade per year of teacher experience.	$\{-2, 0, 2\}$
$\gamma_{\text{cs on te}}$	$\gamma_{\text{cs on te}}$ change in number of years of teacher experience per student	$\{-.1, 0, .1\}$
σ_{te}	Standard deviation of number of years of teacher experience	$\{1\}$

DGP2 Simulation Conditions II

Parameter	Description	Value
σ_{cm}	Between cluster standard deviation	$\{1\}$
γ_{00}	Population mean grade	$\{50\}$
σ_{ϵ}	Error standard deviation.	$\{1\}$

DGP3 Mathematics

$$Grade_{ij} \sim \mathcal{N}(CM_{0j} + \gamma_{sm} \cdot SM_{ij}, \sigma^2_{\epsilon}) \quad (17)$$

$$CM_{0j} \sim \mathcal{N}(\gamma_{00} + \gamma_{cs} \cdot CS_j, \sigma^2_{cm}) \quad (18)$$

$$SM_{ij} \sim \mathcal{N}(SM_j, \sigma^2_{sm}) \quad (19)$$

$$SM_j \sim \mathcal{N}(0 + \gamma_{\text{cs on sm}} \cdot CS_j, \sigma^2_{smg}) \quad (20)$$

$$CS_j \sim \min\{Pois(\lambda), 1\}, \quad (21)$$

$$grade_{ij} \sim N\left(CM_{0j} + \gamma_{sm} SM_{ij}, \sigma^2_{\epsilon}\right) \quad (22)$$

$$CM_{0j} \sim N\left(\gamma_{00}, \sigma^2_{cm}\right) \quad (23)$$

$$grade_{ij} \sim N\left(CM_{0j} + \gamma_{sm} SM_{ij}, \sigma^2_{\epsilon}\right) \quad (24)$$

$$CM_{0j} \sim N\left(\gamma_{00} + \gamma_{cs} CS_j, \sigma^2_{cm}\right). \quad (25)$$

DGP3 Simulation Conditions I

Parameter	Description	Value
Number of Clusters (J)	Number of J clusters for the MCS replication	$\{5, 50, 100\}$
λ	Specifies the Poisson distribution to take one draw of to determine a given cluster's size.	$\{25, 100\}$
γ_{cs}	γ_{cs} point change in mean cluster grade per student.	$\{-.25, 0, .25\}$
γ_{sm}	γ_{te} point change in mean final grade per unit of student motivation.	$\{-2, 0, 2\}$
$\gamma_{\text{cs on sm}}$	$\gamma_{\text{cs on sm}}$ change in units of student motivation per student in cluster j	$\{-.1, 0, .1\}$
σ_{cm}	Between cluster standard deviation of student grades	$\{1\}$

DGP3 Simulation Conditions II

Parameter	Description	Value
σ_{sm}	Within-cluster standard deviation of student motivation	$\{1\}$
σ_{smg}	Between-cluster standard deviation of cluster-mean student motivation	$\{1\}$
γ_{00}	Population mean grade	$\{50\}$

No-effect of between-cluster differences in student motivation. Cluster size creates most (but not all) of the between-cluster differences in student motivation.