Model Validation in the Presence of Multifaceted Error

Challenges for Understanding Teaching Quality with Observation Data

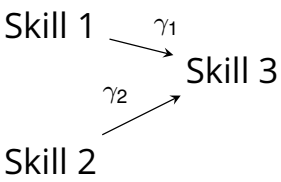
Steffen Erickson

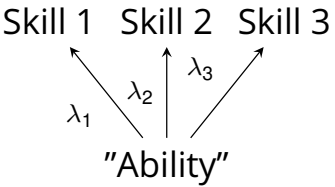
University of Virginia
cns8vg@virginia.edu

Table of Contents

Assessing Performance with Observational Measures

- Interest in assessing performance with observations judged by human raters
- Method used to study teaching quality, where researchers seek to measure a set of skills that are important for high-quality teaching
- Measurements are used to test conceptual models about the relationships between skills





Challenges For Validating Models of Teaching Quality using Observation Data

- **Multifaceted errors define observation conditions**: Large proportion of variance in observation scores results from multifaceted error that define the observation conditions (e.g., raters, lesson conditions, student composition)

(Bartanen & Kwok, 2021; Boguslav & Cohen, 2024; Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012; Mantzicopoulos, French, Patrick, Watson, & Ahn, 2018; Mashburn, Meyer, Allen, & Pianta, 2014)

- **Multifaceted error as omitted variables**: Relationships between skills are confounded by the conditions of the observation (i.e., observers tend to rate individual teachers high (or low) relative to their peers on all the competencies simultaneously (Bartanen, Bell, James, Taylor, & Wyckoff, 2023; Kane & Staiger, 2012))
- **Observation selection bias**: Some teachers systematically receive more favorable observation conditions

Motivating Example

- Interested in the dependence of pedagogical content knowledge (η) on subject matter knowledge (ξ) during mathematics teaching tasks
  - Subject matter knowledge (ξ) - Mathematical understanding and reasoning that enable the "unpacking" of math concepts beyond what is typically needed outside of the classroom (Loewenberg Ball, Thames, & Phelps, 2008)
  - Pedagogical content knowledge (η) - Knowledge of common conceptions and misconceptions about math content that enable teachers to anticipate what students will find confusing or challenging (Loewenberg Ball et al., 2008)
- Measurements of (η) and (ξ) are jointly sampled in conditions defined by the observation error influences of raters (r) and lessons (l)

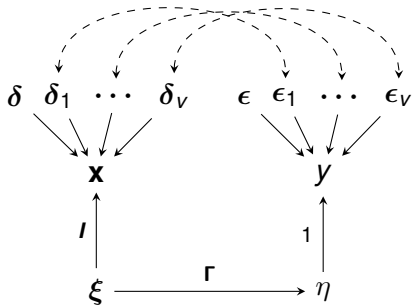
Multifaceted Errors

		p_1	p_2	p_3	p_4
l_1	r_1	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$
	r_2	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$
l_2	r_1	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$
	r_2	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$	$x \longrightarrow y$

- Interested in the true model $\xi \rightarrow \eta$, but we observe the observation-varying model $x \rightarrow y$ depending on the conditions of the measurement procedure
- $\xi \rightarrow \eta$ is a function of $\boldsymbol{\Sigma_p}$
- $x \rightarrow y$ is a function of $\boldsymbol{\Sigma_p}$ **and** $\boldsymbol{\Sigma_l}, \boldsymbol{\Sigma_r}, \boldsymbol{\Sigma_{p \times l}}, \boldsymbol{\Sigma_{p \times r}}, \boldsymbol{\Sigma_{l \times r}}, \boldsymbol{\Sigma_{p \times l \times r}}$

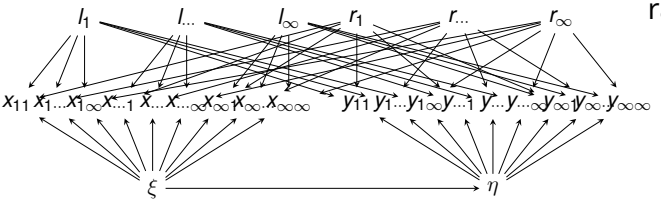
Multifaceted Errors as Confounders

- The bias introduced multifaceted errors is
- $\gamma^* = \gamma[\frac{\phi}{VAR(x)}] + \frac{1}{VAR(x)} \sum_{i=1}^{v} COV(\epsilon_i, \delta_i)$
- The bias depends on
  - $\frac{\phi}{VAR(x)}$, or the reliability of the measurement x
  - and $\frac{1}{VAR(x)} \sum_{i=1}^{v} COV(\epsilon_i, \delta_i)$, the sum of covariances created from taking measurements x and y in conditions defined by the v random effects



Observation Selection Process

Infinite universe of possible measurement conditions

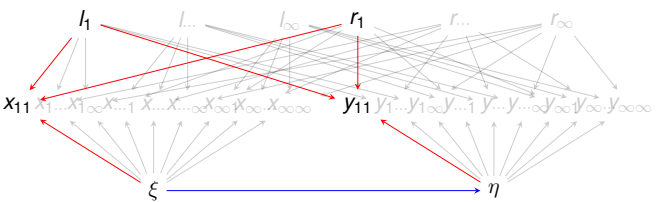


- Measurements of (η) and (ξ) for each teacher are jointly sampled from conditions defined by the ∞ levels of raters and lessons.

		r = 1	r = ⋯	r = ∞
I	1	$x \to y$	$x \to y$	$x \to y$
	⋯	$x \to y$	$x \to y$	$x \to y$
	∞	$x \to y$	$x \to y$	$x \to y$

Observation Selection Process

Sampling procedure



l \ r	1	$\cdots$	∞
1	$x \to y$	$x \to y$	$x \to y$
$\cdots$	$x \to y$	$x \to y$	$x \to y$
∞	$x \to y$	$x \to y$	$x \to y$

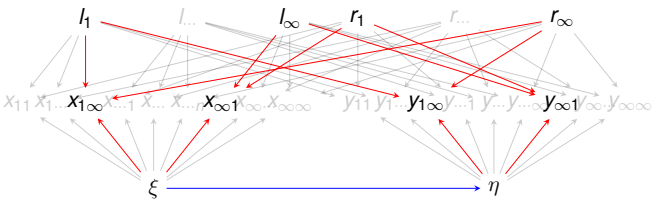
$$x_{11} \leftarrow \xi \to \eta \to y_{11}$$

$$x_{11} \leftarrow l_1 \to y_{11}$$

$$x_{11} \leftarrow r_1 \to y_{11}$$

Observation Selection Process

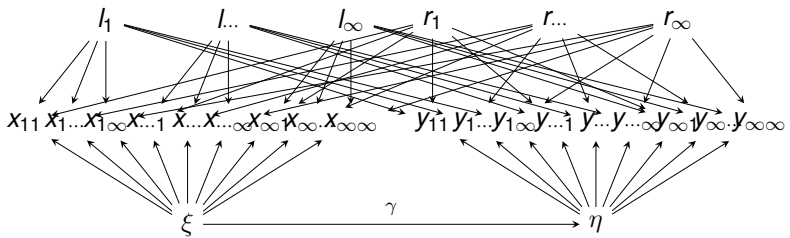
The number of observations and their conditions will vary



		r = 1	r = ⋯	r = ∞
l	1	$x \to y$	$x \to y$	$x \to y$
l	⋯	$x \to y$	$x \to y$	$x \to y$
l	∞	$x \to y$	$x \to y$	$x \to y$

Population Model

- Simulate data-generating model using the General SEM under various conditions to understand the influence of multifaceted errors and observation selection on the model $\xi \rightarrow \eta$



Simulation Design

Simulation Design

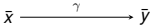
Type	Factors	Conditions
Data Generation Factors	ξ and η loadings	$\times$ .5, $\times$ 1, $\times$ 1.5, $\times$ 2, $\times$ 3
	Rater & lesson factor loadings	$\times$ .5, $\times$ 1, $\times$ 1.5
	ξ and η factors variances	$\times$ .5, $\times$ 1, $\times$ 1.5
	Rater & lesson factors variances	$\times$ .5, $\times$ 1, $\times$ 1.5
	Rater & lesson mean deviations	$\times$ 1, $\times$ 2, $\times$ 3
	Sample	100, 500, 1000
	Rater & lesson levels	4 & 8
Sampling Factors	**Random sample of measures**	yes, no
	Sampling design	**PbyLbyR**, PbyLinR, PbyRinL, **LinRinP**
	Measurements per Teacher	1, 2, 3, 4, 6, 8, 16, 32

Notes: All factors fully crossed in the simulation. Data generation conditions edited fixed parameters. Average starting factor loadings = .75, average starting error variances = .75, average starting factor variance = 1.5, average starting rater & lesson mean deviations = -1 to 1.

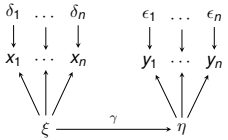
method

Results

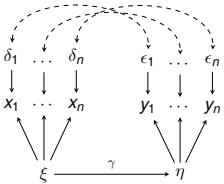
1. Random error vs. observation confounding with Model 1
2. Adjusting for random error and observation confounding with Model 2 & Model 3
3. Impact of non-random sampling on correctly specified model with Model 3



Model 1: Average Score Estimates

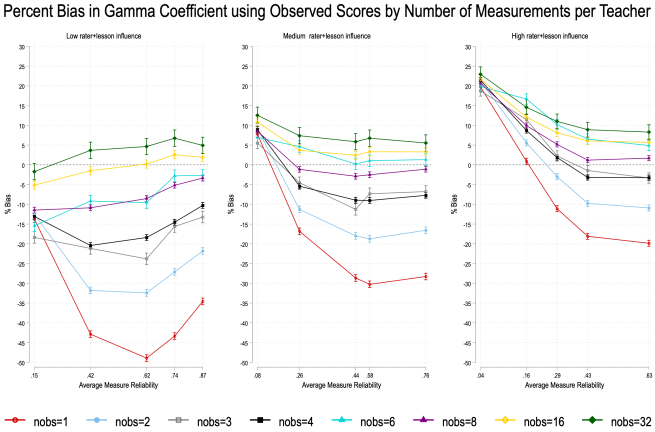


Model 2: Measurement Error Correction

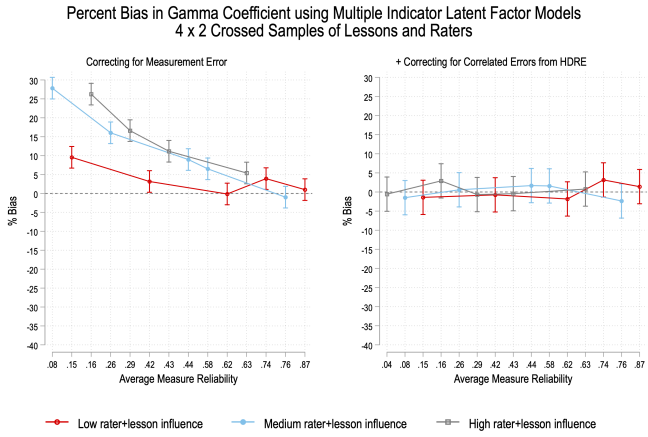


Model 3: Correlated Errors Across Observations

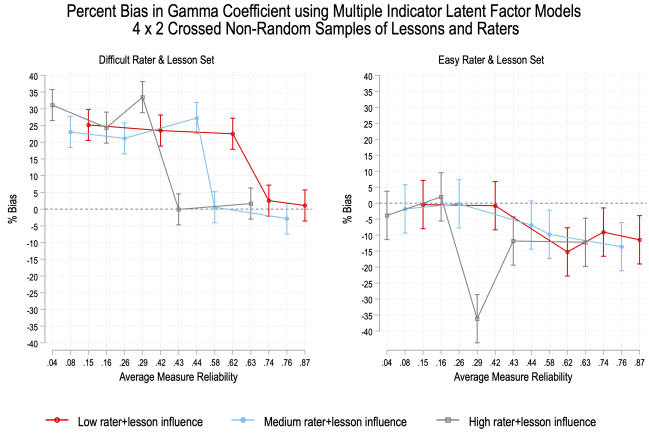
1. Random Error vs. Observation Confounding



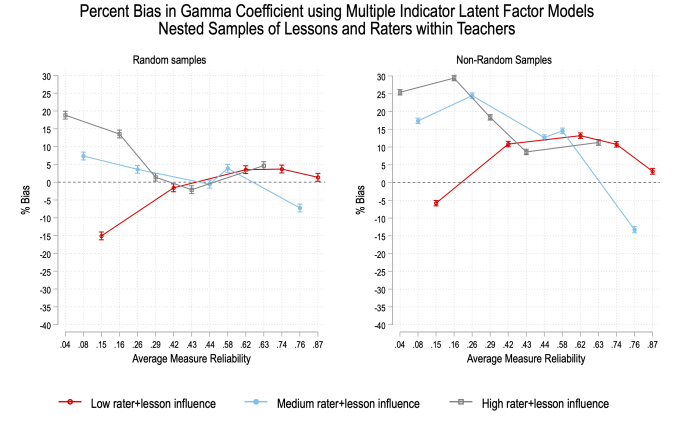
2. Adjusting for Random Error and Observation Confounding



3. Impact of Non-Random Sampling on Correctly Specified Model



3. Impact of Non-Random Sampling on Correctly Specified Model



Implications

- Jointly sampling measurements in the same observation conditions confounds the relationships between skills
- Confounding and random error make the bias in estimates difficult to predict
- Flexible SEMs are not sufficient in the presence of non-random observation selection

Next Steps

- Need for systematic and efficient designs to collect observations of teachers
  - Planned missing data with optimal designs for sub-samples
- Need for
  - Efficient estimation methods with clear assumptions
  - Modeling the selection process of observations

Thanks for your Time!

Thanks for your time!
Contact: **cns8vg@virginia.edu**
Simulating SEMs with Multifaceted Errors:
https://github.com/steffenerickson/latentvariablesunderHDRE

Simulation Method

- Start with the General Model

$$\eta = \alpha + \boldsymbol{B}\eta + \boldsymbol{\Gamma}\xi + \zeta$$

$$\boldsymbol{y} = \upsilon_{\boldsymbol{y}} + \boldsymbol{\Lambda_y}\eta + \epsilon$$

$$\boldsymbol{x} = \upsilon_{\boldsymbol{x}} + \boldsymbol{\Lambda_x}\xi + \delta$$

- Construct the model implied population covariance and mean vector

$$\boldsymbol{\Sigma}(\theta) = \begin{bmatrix} \boldsymbol{\Lambda_y}(\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi})\left[(\boldsymbol{I}-\boldsymbol{B})^{-1}\right]'\boldsymbol{\Lambda_y}' + \boldsymbol{\Theta_\epsilon} & \boldsymbol{\Lambda_y}(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma\Phi\Lambda_x}' \\ \boldsymbol{\Lambda_x\Phi\Gamma}'\left[(\boldsymbol{I}-\boldsymbol{B})^{-1}\right]'\boldsymbol{\Lambda_y}' + \boldsymbol{\Theta_\epsilon} & \boldsymbol{\Lambda_x\Phi\Lambda_x}' + \boldsymbol{\Theta_\delta} \end{bmatrix}$$

$$\upsilon = \begin{bmatrix} \boldsymbol{\Lambda_y}(\boldsymbol{I}-\boldsymbol{B})^{-1}(\alpha + \boldsymbol{\Gamma}\kappa) \\ \boldsymbol{\Lambda_x}\kappa \end{bmatrix}$$

Return to presentation

Simulation Method

- Using loading matrices $\boldsymbol{\Lambda_y}$ and $\boldsymbol{\Lambda_x}$ as design matrices, I can construct factorial designs to sample measures of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ from
- For example,

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ \lambda_{21} & 0 & \lambda_{23} & 0 & 0 & 1 \\ \lambda_{31} & 0 & 0 & 1 & \lambda_{35} & 0 \\ \lambda_{41} & 0 & 0 & \lambda_{44} & 0 & \lambda_{46} \\ 0 & 1 & \lambda_{53} & 0 & \lambda_{55} & 0 \\ 0 & \lambda_{62} & \lambda_{63} & 0 & 0 & \lambda_{66} \\ 0 & \lambda_{72} & 0 & \lambda_{74} & \lambda_{75} & 0 \\ 0 & \lambda_{82} & 0 & \lambda_{84} & 0 & \lambda_{86} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ l_1 \\ l_2 \\ r_1 \\ r_2 \end{bmatrix}$$

Return to presentation

Simulation Method

- I construct data in the following manner
- Let the model implied covariance matrix $\mathbf{\Sigma}(\theta) = \mathbf{A}'\mathbf{A}$
- $\mathbf{A} = \mathbf{U}\sqrt{\mathbf{D}}$ from $\mathbf{\Sigma}(\theta) = \mathbf{U}'\mathbf{D}\mathbf{U}$
- $\mathbf{X}$ is first generated as $\mathbf{X} \sim N(0, \mathbf{I})$.
- $\mathbf{Y} = \mathbf{A}'\mathbf{X} + \boldsymbol{\upsilon}$, with $\mathbf{Y} \sim N(\boldsymbol{\upsilon}, \mathbf{\Sigma}(\boldsymbol{\theta}))$.
- Random samples are pulled from $\mathbf{Y}$
- Example Code

Return to presentation

References I

Bartanen, B., Bell, C., James, J., Taylor, E. S., & Wyckoff, J. H. (2023). " refining" our understanding of early career teacher skill development: Evidence from classroom observations. edworkingpaper no. 23-845. *Annenberg Institute for School Reform at Brown University*.

Bartanen, B., & Kwok, A. (2021). Examining clinical teaching observation scores as a measure of preservice teacher quality. *American Educational Research Journal*, *58*(5), 887–920.

Boguslav, A., & Cohen, J. (2024). Different Methods for Assessing Preservice Teachers' Instruction: Why Measures Matter. *Journal of Teacher Education*, *75*(2), 168–185. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

References II

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. (Publisher: ERIC)

Loewenberg Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of teacher education*, *59*(5), 389–407.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, *23*(1), 24–46.

References III

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400–422.