Approximate Measurement Invariance: Measuring Classroom Teaching Quality

Ben Kelcey
Dan McGinn
Heather Hill

ben.kelcey@gmail.com

Project Team

- National Center for Teacher Effectiveness
  - gse.harvard.edu/ncte
  - PIs: Tom Kane, Heather Hill & Doug Staiger
  - Dan McGinn, Matthew Kraft, Mark Chin, Charalambos Charalambous

Context

- Recent policy has charged districts to differentiate among teachers in terms of their effectiveness
- Value-added indices are one approach
  - They tell us who is more effective, but not why and how we might improve teaching
- A complementary approach to value-added measures is classroom observations
  - The task of measuring teaching through observations challenges decision makers to make precise their criteria for effective teaching by anchoring effective teaching in specific and observable criteria

Reality

- The potential of classroom observations is mitigated by features of the observational environment and system used to conduct observations, e.g.,
  - Fallible indicators
  - Rater differences (e.g., rater severity)
  - Atypical observations
- These features potentially confound construct-irrelevant variance with teaching quality

Rater Differences

- Rater difference manifest in many forms
- Some commonly noted forms (e.g., Eckes, 2009):
  - Different understandings of rating scale categories
    - Undue emphasis on particular features, e.g., halo
  - Different use of rating scale categories
    - Differences in severity/leniency of assigned ratings
    - Central (or extreme) tendency
    - Rater drift

Why Are Rater Differences a Problem?

- Rater differences:
  - Introduce construct-irrelevant variance
  - Introduce variability in the structure of the scale
  - Score categories for a competency may no longer have a consistent meaning across raters
- Collectively, rater differences have the potential to undermine the reliability and validity of assessments

Common Design-based Approaches

- Design-based approaches to minimizing rater differences
  - Have extensive training and practice for raters
  - Certification tests
  - Ongoing monitoring
  - Calibration
  - Retraining

Prior Research

- Rater variability is substantial even among well trained and experienced raters who are monitored and pass regular calibration tests
  - e.g. MET study

Variance decomposition for CLASS (overall)

Component	Variance
Teacher	0.31
Observation	0.27
Rater	0.08
Residual	0.34

Analytic Treatments of Rater Differences

- Literature suggests design-based approaches are generally not sufficient for addressing differences among raters
  - e.g., Reliability is necessary but not sufficient
- Some analytic-based approaches
  - Generalizability theory
  - Hierarchical rater model
  - Many-facet Rasch model

Measurement Invariance Across Raters

- Extant measurement models fundamentally assume measurement invariance across raters
  - In other words, raters apply the indicators to measure latent teacher quality in the identical way
- Invariance is a prerequisite for meaningful comparisons…
- Yet, there is clear counter evidence in most assessments of measurement non-invariance

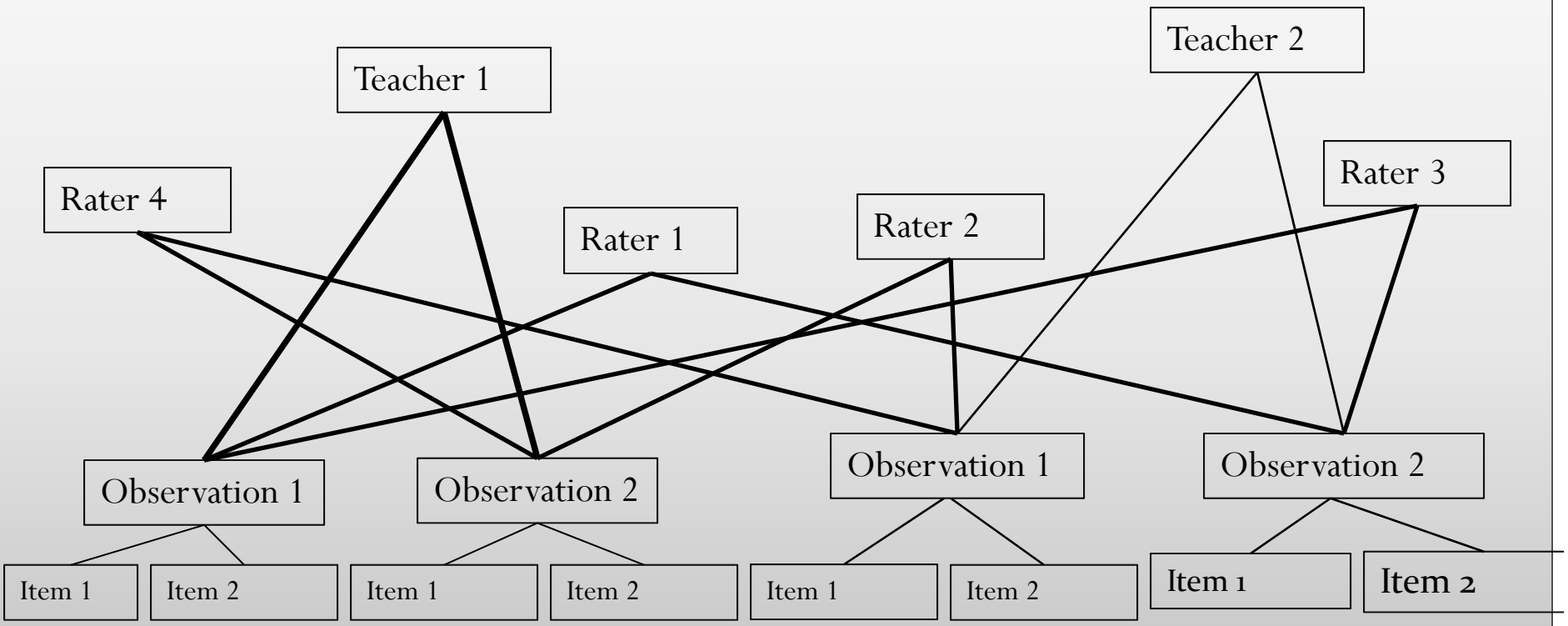
A Rock and a Hard Place

- We want to assess classroom teaching quality but we know even with extensive training raters are likely to differ in complex ways
- What to do?
- Approximate measurement invariant models
  - Examine the extent to which an approximate measurement invariant item response model outperforms models that assume invariance

Design of Study

- 250 grade four and five mathematics teachers across five districts
- For each teacher
  - Teaching was evaluated over multiple observations
  - Teaching was evaluated using the Mathematical Quality of Instruction (MQI) Protocol
    - E.g., Learning Mathematics for Teaching. (2011). Measuring the Mathematical Quality of Mathematics teaching. *Journal for Mathematics Teacher Education* 14(1), 25-47.
  - Each observation was evaluated by two (of 39) raters
  - Value-added scores for teachers

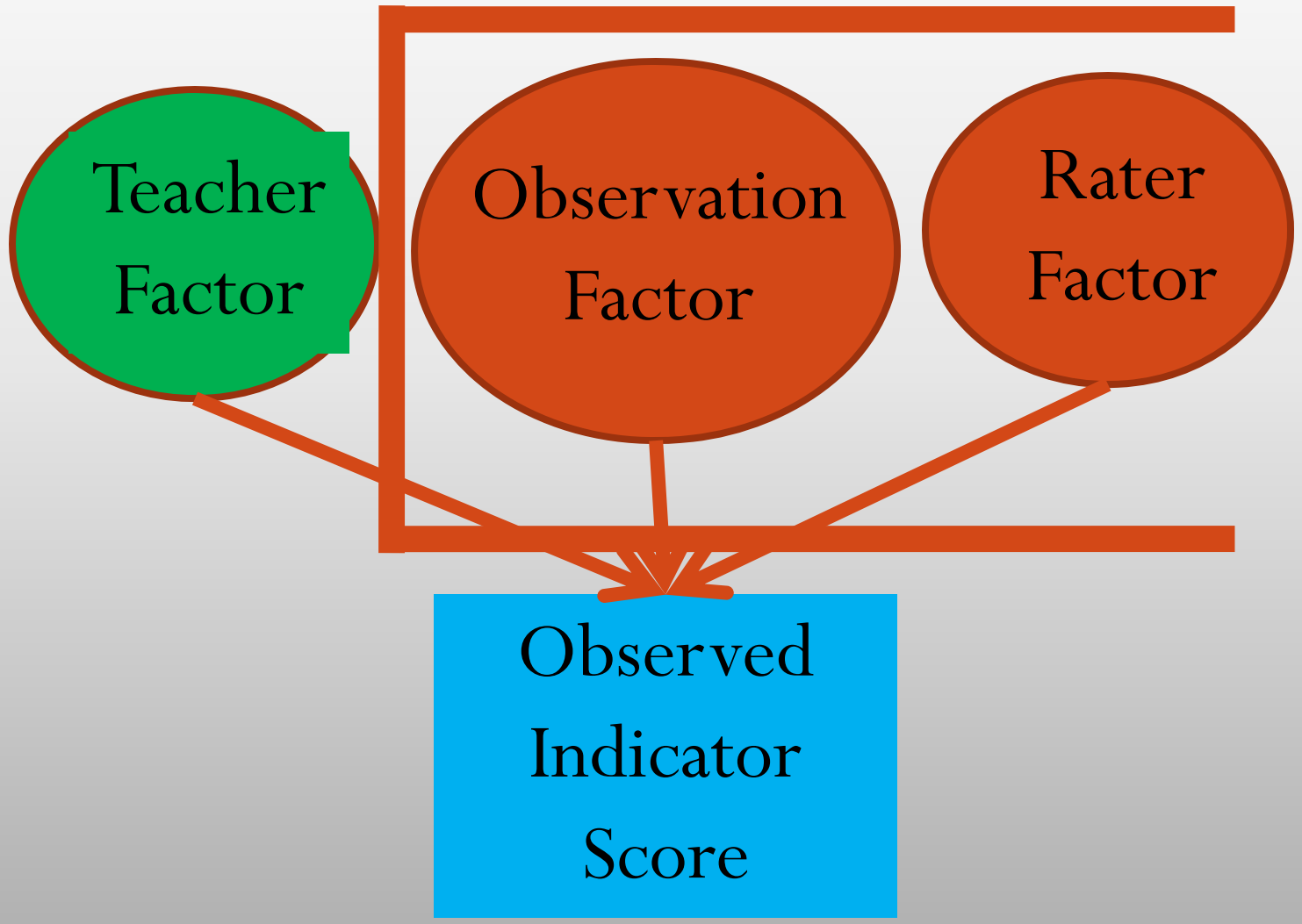
Structure of Observed Data



General Teaching Quality

- For today's presentation we focus on a single general dimension of teaching quality
- Indicators (items) of this quality are:
  - Richness of mathematics
  - Working with students and mathematics
  - Errors and imprecision
  - Student participation in meaning-making and reasoning
  - Whole-class discussion
  - Classroom work is connected to mathematics

Signal v. Noise



Item Response Theory

- Item Response Theory (IRT)
  - Describes teaching observation ratings as a function of an underlying and unobserved latent trait
  - Data analyzed at item level such that the probability of receiving a high quality rating is a function of
    - underlying and unobserved latent trait (θ)
    - item parameters
  - Potentially separates teaching quality from construct-irrelevant variance

Item Response (IRT)

- Graded Response Model

$$P(Y_{iotr} = k) = P(Y_{iotr} \geq k-1) - P(Y_{iotr} \geq k) =$$

$$\Psi(a_i(\theta_t - d_i^{k-1})) - \Psi(a_i(\theta_t - d_i^{k}))$$

- Y_{iotr} is the score for item i in observation o for teacher t rated by rater r
- a_i represents the discrimination parameter for item i
- θ_t represents teacher t's level of teacher quality
- K is the number of categories items are graded on (three) with k as a specific category
- $d_i^{(1)},\ldots,d_i^{(K-1)}$ are a set of K-1 ordered item difficulty intercepts.

IRT

- Some implications of the IRT model:
  - Local independence among ratings within the same observation
    - Conditional upon teaching quality (theta), ratings within the same observation were independent
    - We made no provisions for dependencies among ratings in the same observation
    - What if the teacher had a bad day or was observed during an atypical lesson? Chances are, if they are having a bad day, they are more likely to do poorly on many indicators (local dependence among indicators beyond that induced by teaching quality)

Multilevel Item Response Theory (MLIRT)

- Multilevel graded response model (Ψ is the logistic CDF):

$$P(Y_{iotr} = k) = P(Y_{iotr} \geq k-1) - P(Y_{iotr} \geq k) =$$

$$\Psi(a_i(\theta_t + \alpha_{otr} - d_i^{k-1})) - \Psi(a_i(\theta_t + \alpha_{otr} - d_i^{k}))$$

- Y_{iotr} is the score for item i in observation o for teacher t rated by rater r,
- a_i represents the discrimination parameter for item i,
- θ_t represents teacher t's stable level of teacher quality
- α_{otr} is the deviation specific to observation o for teacher t rated by rater r
- K is the number of categories items are graded on (three) with k as a specific category
- $d_i^{(1)},\ldots,d_i^{(K-1)}$ are a set of K-1 ordered item difficulty intercepts.

MLIRT

- Some implications of the MLIRT model:
  - All raters are identical
    - We made no provisions for the fact that different raters rated different observations
- Local independence among observations rated by the same raters
  - Conditional upon teaching quality (theta) and the dependence of scores within the same observation (alpha), ratings were independent
    - We made no provisions for dependencies among ratings given by the same rater

Cross-class Multilevel Item Response Theory (CCIRT)

- Multilevel graded response model (Φ is the normal CDF):

$$P(Y_{iotr} = k) = P(Y_{iotr} \geq k-1) - P(Y_{iotr} \geq k) =$$

$$\Phi(a_i(\theta_t + \gamma_r + \alpha_{otr} - d_i^{k-1})) - \Phi(a_i(\theta_t + \gamma_r + \alpha_{otr} - d_i^{k}))$$

- Y_{iotr} is the score for item i in observation o for teacher t rated by rater r,
- a_i represents the discrimination parameter for item i,
- θ_t represents teacher t's stable level of teacher quality
- γ_r is a fixed effect for rater r's level of leniency
- α_{otr} is the deviation specific to observation o for teacher t rated by rater r
- K is the number of categories items are graded on (three) with k as a specific category
- $d_i^{(1)},\ldots,d_i^{(K-1)}$ are a set of K-1 ordered item difficulty intercepts.

CCIRT

- Some implications of the CCIRT model:
  - Raters are allowed to be different…but, only up to a simple shift in severity
  - Rater severity effects suggest that raters are uniformly more severe or lenient on all indicators—what if a rater is more severe on one indicator but more lenient on another
  - Further, what if raters vary in their ability to differentiate among quality levels…and what if this ability varies by indicator
- Are we really placing teachers really on the same scale?

Approximate Invariance

- Invariance is a prerequisite for meaningful comparisons
  - How can we compare teachers who were rated using different rulers?
- **Approximate Measurement Invariance**
  - Differences in raters' rulers are conceived as random measurement deviations
  - Rather than assume invariance, we acknowledge and adjust for differences as best we can
  - Random item effects capture the measurement deviation due to rater differences/differential rater functioning

Cross-class Multilevel Item Response Theory with Random Item Effects (CCIRT-RIE)

- Graded response model with random loadings and thresholds (Φ is the normal CDF):

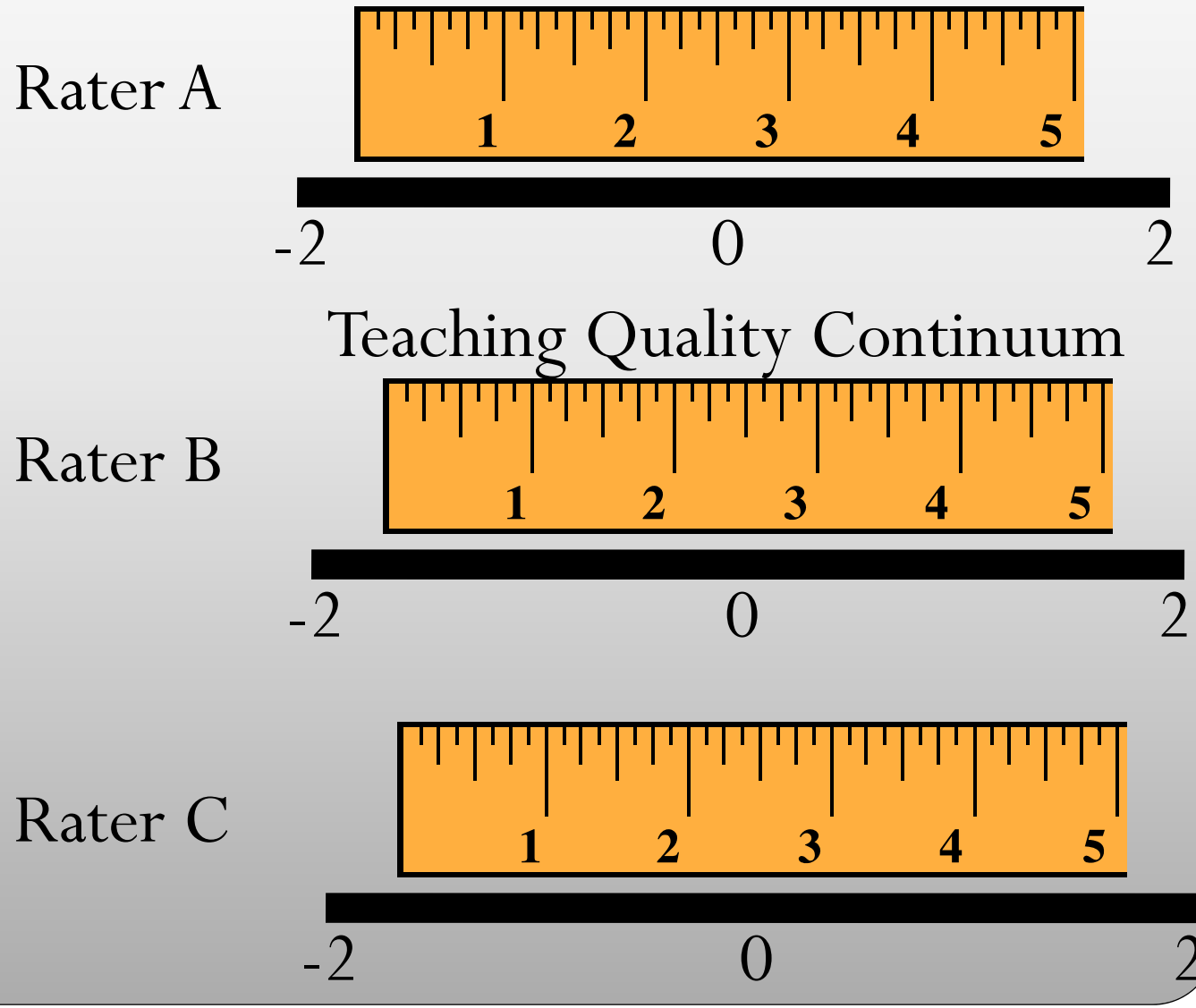
$$P(Y_{iotr} = k) = \Phi(a_i\theta_t + a_{ir}\alpha_{osr} + a_{ir}\gamma_r - d_{ir}^{k-1}) - \Phi(a_i\theta_s + a_{ir}\alpha_{osr} + a_{ir}\gamma_r - d_{ir}^{k})$$

$$a_{ir} \sim N(a_i, \sigma_{a,i}),\ d_{ir}^k \sim N(d_i^k, \sigma_{d,i}),$$

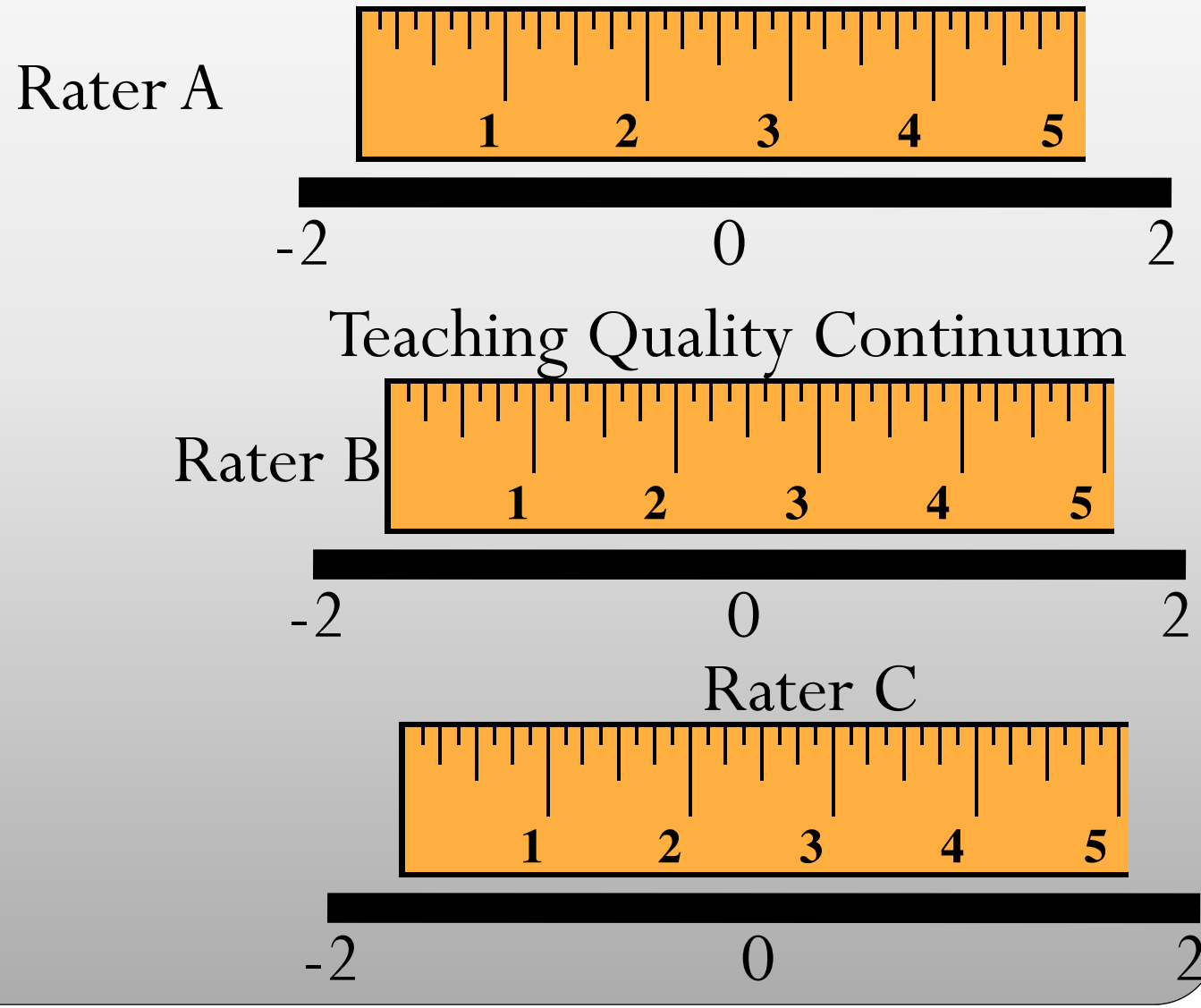
$$\alpha_{otr} \sim N(0,1), \gamma_r \sim N(0, \sigma_{\gamma_r}), \theta_t \sim N(0, \sigma_{\theta_t})$$

- Both discrimination (a) and difficulty thresholds (d) vary across raters to compensate for invariance

MLIRT with no rater effects

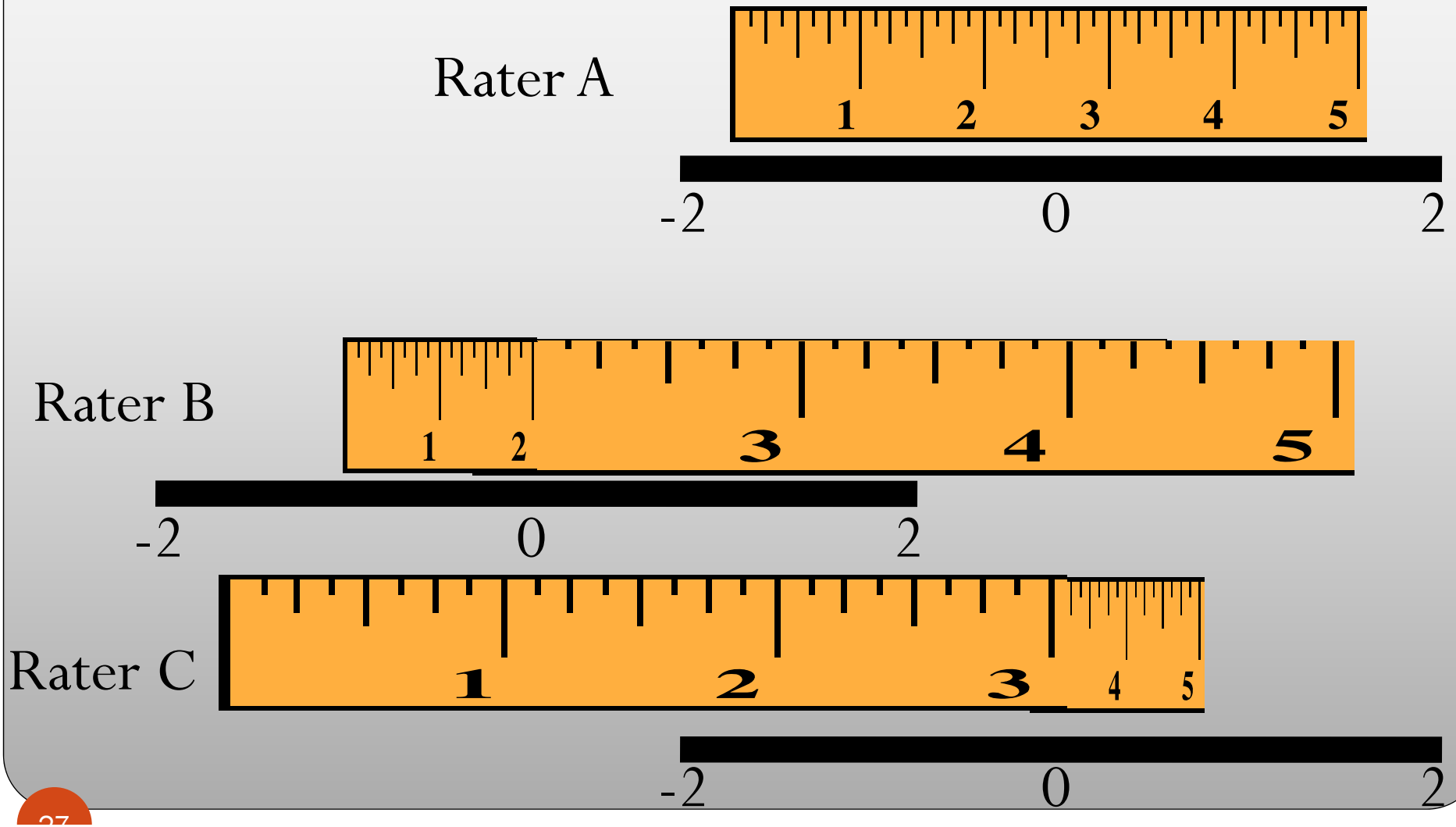


CCIRT with rater severity adjustment



CCIRT-RIE

- Perhaps raters' rulers don't just shift, they have different thresholds for different indicators



Do Indicator Properties Vary by Rater?

Variance of Random Item Effects

Quality Indicators	Discrim	Difficulty
Richness of mathematics	0.11	0.42
Working with students and mathematics	0.14	0.80
Errors and imprecision	0.10	1.00
Student participation in meaning-making and reasoning	0.18	0.84
Whole-class discussion	0.03	1.01
Classroom work is connected to mathematics	0.09	0.21

External Validation

- Do our adjustments really reduce the noise introduced by construct-irrelevant sources of variation (e.g., by differences in rater severity)?
- Do our adjustments address construct-irrelevant variance in ways that relate to achievement and teacher effectiveness?
- How do scores from these models relate to teacher value-added scores?
- Do all adjustments improve our results?

Preliminary Evidence

- Correlation of classroom observation and VA scores
  - 55% improvement
  - Incorrect adjustments?
- Correlation triples
- Approach shows promise, but results are still very preliminary

	Coef (SE)	*t*
Simple averages	0.07(0.08)	0.9
MLIRT	0.11(0.07)	1.6
CCIRT	0.09(0.07)	1.3
CCIRT-RIE	**0.20***(0.07)	2.7

Implications

- Adjustment for rater differences is a double-edged sword
  - On one hand it adjusts for rater differences in specific ways
  - On the other hand, you must assume (or verify) that the adjustments are appropriate
- However, without adjustments it is very difficult to make valid comparisons
- Approximate invariance provides a rich and flexible alternative to more rigid models that assume invariance because it attempts to acknowledge and compensate for invariance among raters

Thanks

Ben Kelcey

ben.kelcey@gmail.com

University of Cincinnati