Supporting States in Enhancing Reliability of QRIS Ratings

What we are learning from CLASS?

Agenda

- Possible consequences of biased data within the QRIS framework
- Factors to improve reliability of data
- Next steps

Challenges & Opportunities

- **Defining quality**– paying attention to what matters
- **Ensuring fairness** – measuring quality well
- **Improving practice**
  - Providing resources to support teachers to be more effective
  - Using observational data systematically to inform decision making

What if we get it wrong?

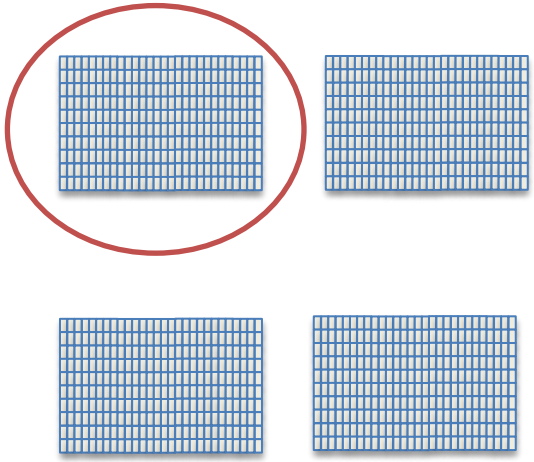
- **Defining effectiveness** – paying attention to things that don't matter OR failing to pay attention to things that do matter

- **Ensuring fairness** – measuring quality poorly leading to inaccurate decision making and feedback AND lawsuits

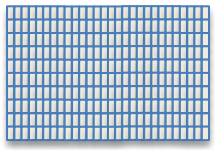
- **Improving practice –** focusing so much on evaluation that we forget to attend to and invest in supporting teachers and programs

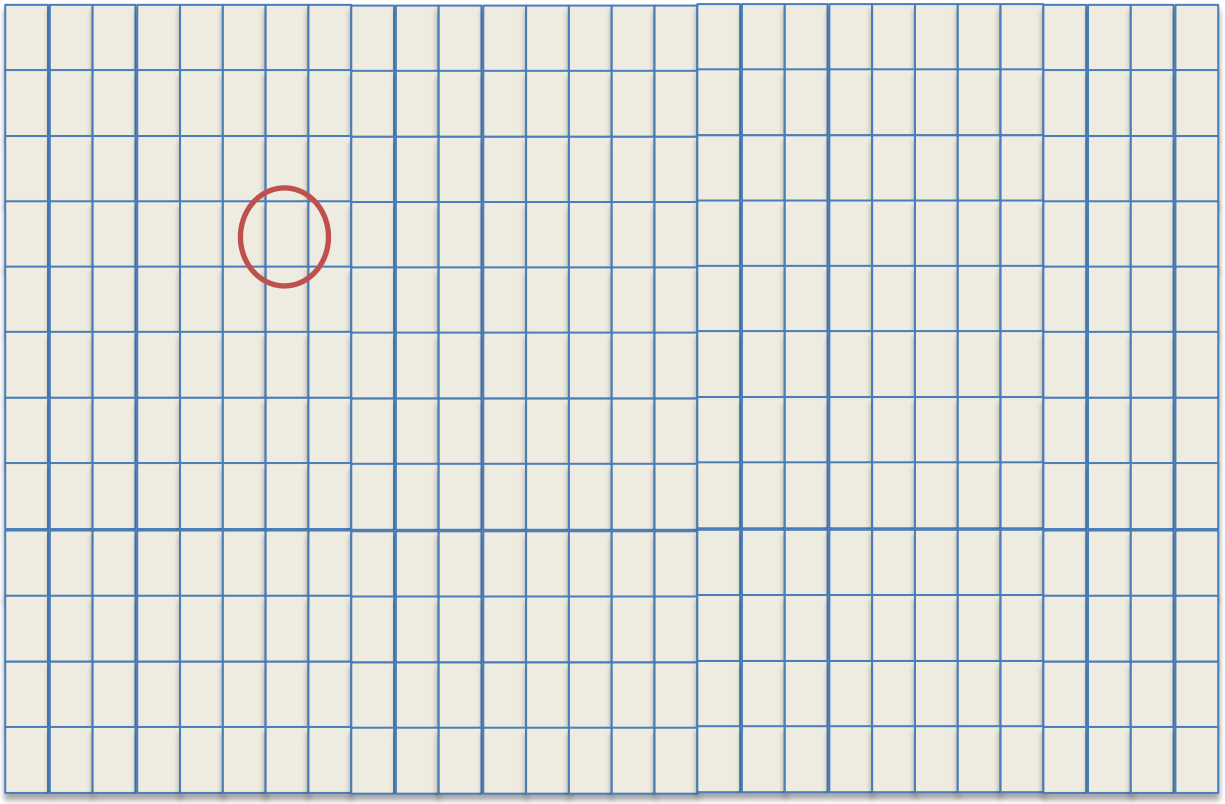
Data Collection: An Illustration…

A child care center has 4 classrooms – we observe in 1

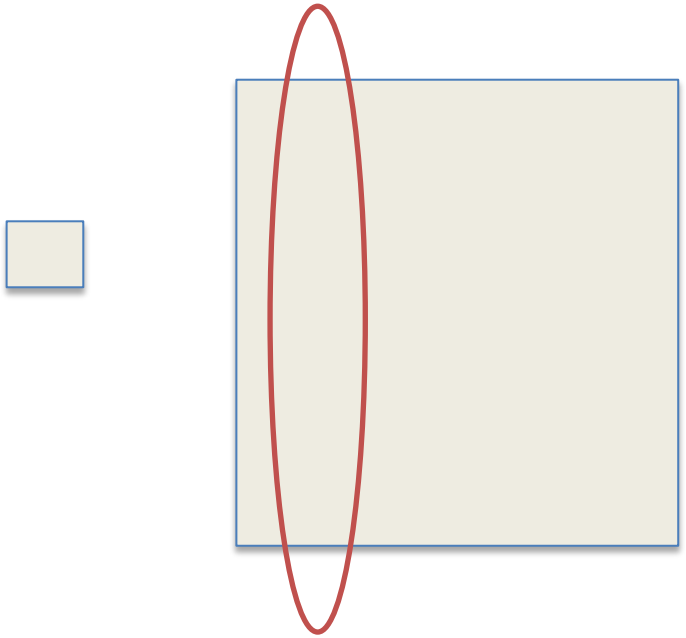


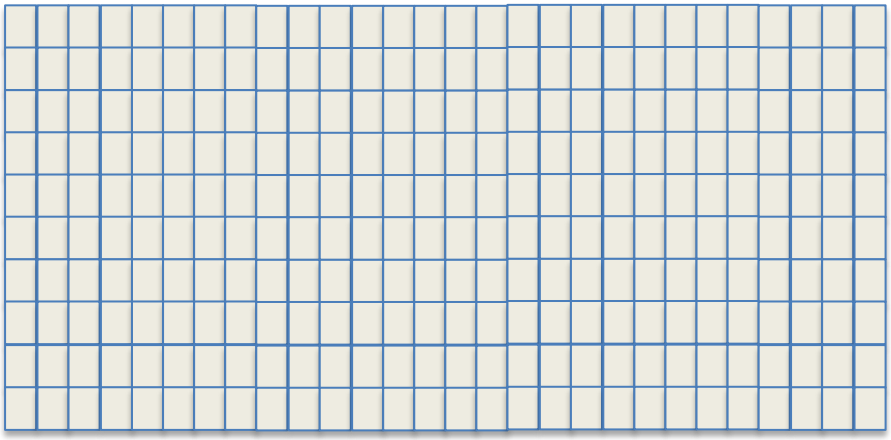
Each classroom has children for about 340 days a year – we observe one day

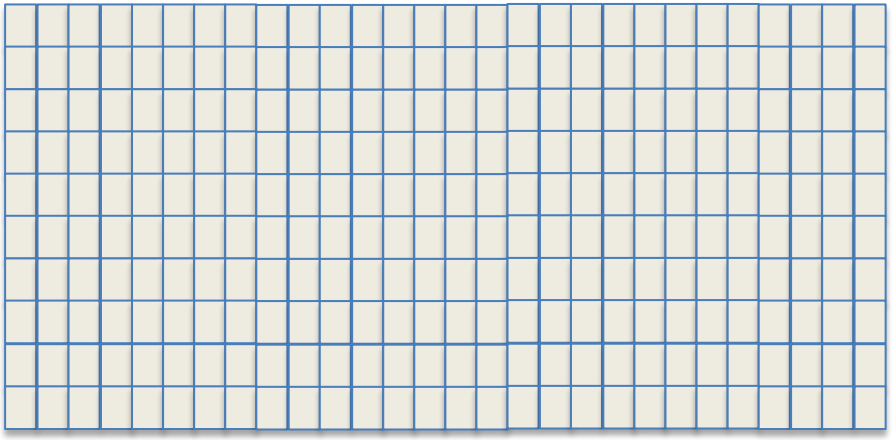


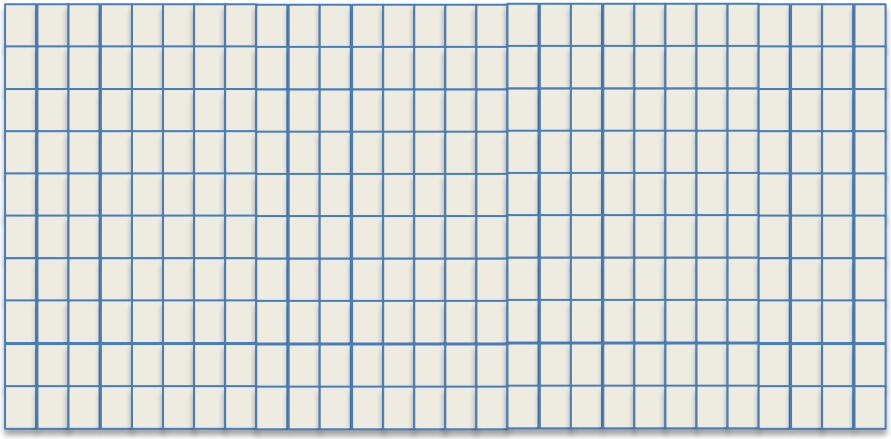


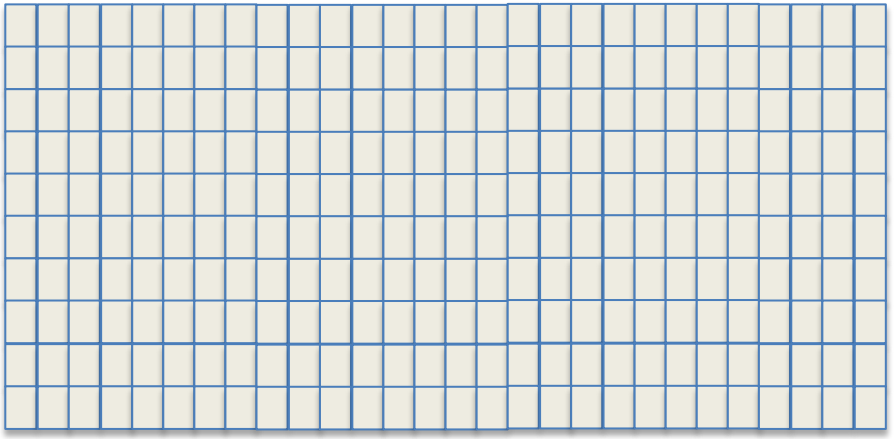
Each day children are in the classroom for 9 hours – we observe for 2











Ensuring fairness

- If we are to use this one estimate, we need to ensure that it is the least biased possible estimate.
- We need more systematic study of error **(systematic bias)** in ratings so that we can provide practical recommendations to the field.

- Systematic bias: a persistent error that cannot be attributed to chance

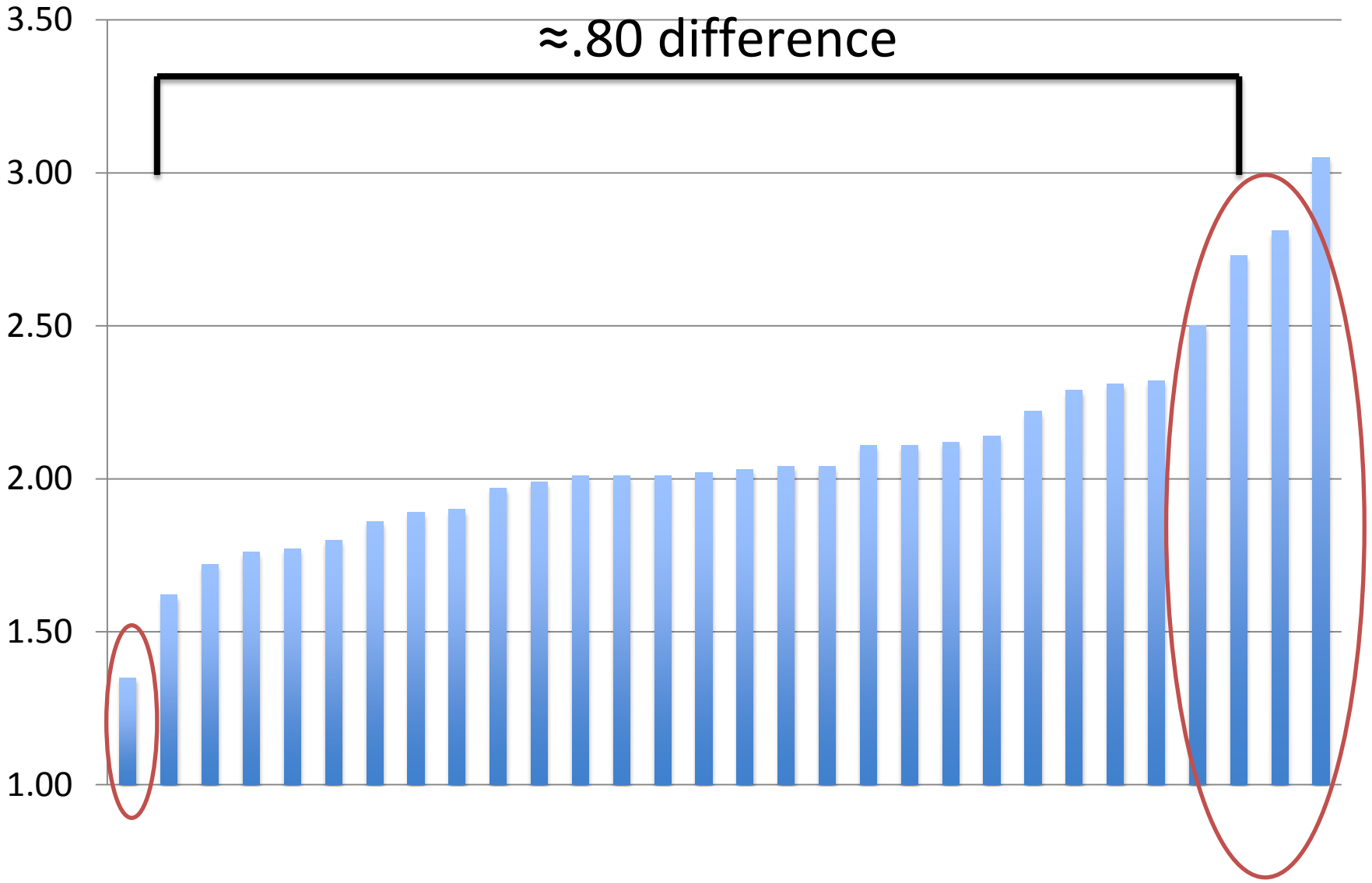
Facts

- There is more variance between classrooms than within classrooms in a given day – (but need updated data on this)
  - ➥Observe in more classrooms, rather than for a longer time in each classroom
- Raters are biased
- Good training can reduce bias
- But bias remains in almost anything “human scored”

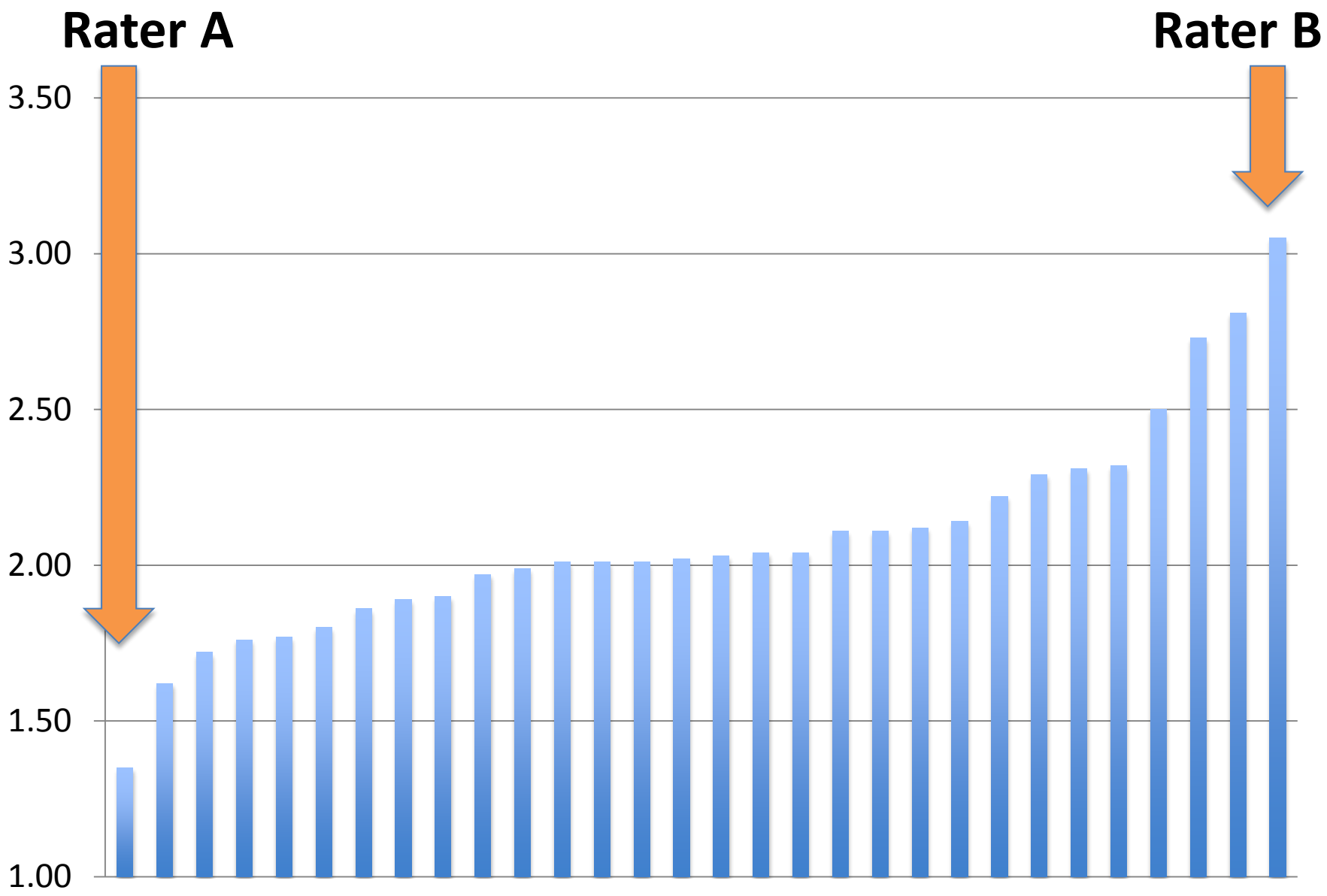
An illustration: Bias

- Data from two large CLASS data collections – not research studies
- All raters passed initial CLASS reliability test
- In both cases there was ongoing calibration (monthly, quarterly)

Average Instructional Support Scores by Rater



Example



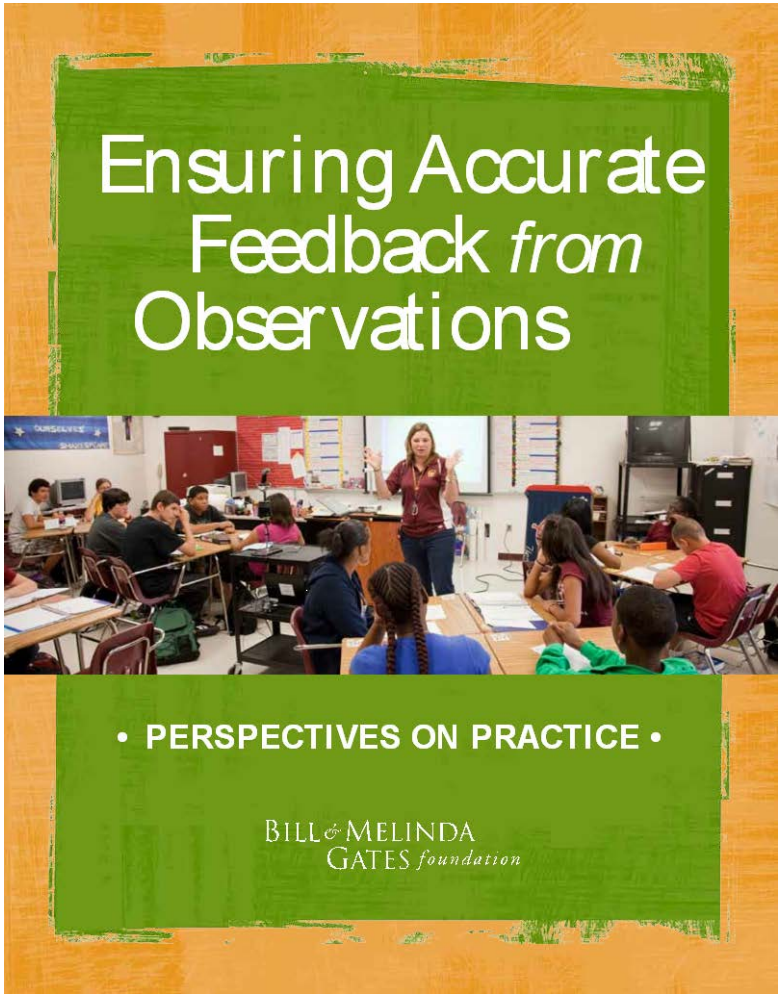
A real life example

- If Rater A scores all classrooms in the Red Program
- And if Rater B scores all classroom in the Green Program
- What will happen to the scores at the program level?

- For QRIS, the systematic bias in the raters could lead to lower subsidy vouchers.

What can we do?

1. Select people with less bias
2. Train them well
3. Provide ongoing support
4. Send out teams of raters



Resource –
http://www.gatesfoundation.org/
http://www.teachstone.org

1. Selection of Participants

- Factors related to levels of reliability on the CLASS test:
  - Cash et al., 2012
    - Individual beliefs about children and teaching
    - Group level of beliefs about children
  - Internal Teachstone study of 460 training participants
    - Job category – teacher and administrators less likely to pass than researchers

2. Training

- 2 day training for CLASS is a minimum – what would happen if we did additional training?
- 80% within 1 is the certification cut-off – but should we use a more rigorous level?
- Who trains? Need to know more about what makes trainers effective.
- Randomized trials of training options?

3. Calibration

- Frequency and intensity of calibration
- Consequences of not scoring accurately?
- Regular review of data as it comes it (e.g. double coded data)

4. Teams

- Send out teams of raters to observe in centers, rather than assigning a single rater to a center

- If you need more than one team of observers, make sure to mix up the teams as they go to different centers

What we don't know

- Sampling classrooms within programs---how many?

- How much of a boost in reliability we get for adding each observer, day, etc – cost effectiveness?

Summary

- To increase reliability of CLASS scores
  - Know that individual characteristics influence the likelihood of passing
  - Adhere to at least the 2 day training guidelines
  - Ongoing review calibration data to identify problems areas
  - Send out teams of raters in high stakes scoring to reduce systematic bias
- More work to be done

Thanks!

UNIVERSITY of VIRGINIA
CASTL
CURRY SCHOOL OF EDUCATION

Bridget Hatfield, bhatfield@virginia.edu