

Modernizing Measurement of Early Childhood Classroom Quality

*Collaborative and
Open Source Strategies for
Continuous Measure
Improvement*

THE
UNIVERSITY OF
ILLINOIS
AT
CHICAGO



Presentation at the
2019 Symposium on Children
Quality in Early Learning Environments
Crane Center for Early Childhood Research & Policy
Ohio State University

by Dr. Rachel Gordon

Professor and Director of Research Training
Department of Sociology
University of Illinois at Chicago

Fellow and Associate Director for the Social Sciences,
Institute for Health Research and Policy
Senior Scholar and Chair of Education and Learning Working Group,
Institute of Government and Public Affairs



A Preview



Preview

- What is the **state of the evidence** for widely used observational measures of early childhood classroom quality?
- What are **focal concerns** about these measures, and key **challenges** for observational measures of classroom quality?
- How can **modern strategies for measurement** help to address these concerns and challenges?



Preview: Three Focal Issues

- **Rater Effects**
- **Item Variation**
- **Standard Error of Measurement**



Preview: First Focal Issue

- **Rater Effects:**

- Do different raters score similarly to each other (*inter-rater reliability*)?
- Does a rater score similarly across classrooms and time (*intra-rater reliability*)?
- *To the extent that inter- and intra-rater reliability are low, uses that rely on a single rater offer an uncertain (and some would say “unfair”) signal.*
- *A classroom’s score will depend, in part, on whether the classroom happens to be rated by a harsher or more lenient rater.*



Preview: Second Focal Issue

- **Item Variation:**

- Do item scores vary meaningfully across classrooms?
- Does item content well reflect the full spectrum of the construct?
- *To the extent that item variation is limited we may have traditional ceiling/floor effects.*
- *We will be less able to distinguish between classrooms, less able to detect growth in quality (due to limited variation on “Y”), and less able to predict change in children’s outcomes (due to limited variation on “X”).*



Preview: Third Focal Issue

- **Standard Error of Measurement:**

- Do we have a strong signal (precise estimate) of the classroom's level of quality?
- Or, is the range of best guesses of the classroom's quality level quite wide?
- *To the extent that our range of best guesses is quite wide, we will be less able to detect differences among classrooms (since our range of guesses about them will overlap).*
- *And, we will be less able to predict how classroom quality grows over time or how classroom quality predicts children's growth (since there will be more "unexplainable" variation--more noise).*



Preview: The Importance of Considering Evidence in Relation to Each Proposed Use

- **“High Stakes” Policy Use**
- **Research Use**
- **Practice Use**



Preview: High Stakes Use

- **High stakes use** of measures of early childhood classroom quality **grew** over recent decades.
 - **Head Start** opted for a version of the Classroom Assessment Scoring System (CLASS) designed for preschool classrooms (***CLASS PreK***; Pianta, La Paro, & Hamre, 2008) in its **Designation Renewal System** (DRS; Public Law 110-134).
 - A 2017 state scan found that the ***CLASS PreK*** and the ***Environment Rating Scales*** (2019) were most often used in the state **Quality Rating and Improvement Systems (QRIS)** that incorporated observational tools (QRIS Compendium, 2017).
 - Although details vary across policies, most **compare some aggregation of scores** initially made by a single rater based on a couple hours of observation on a single day to **specific cutoff values**.



Preview: Limited Evidence

- Concerns about the **limited evidence for these high stakes uses** have grown.
 - **Mashburn (2017)** careful enumeration of assumptions of Head Start DRS:
 - Majority of **variance** in CLASS scores attributable to conditions of observation, including **who conducted the observation**.
 - Little evidence supporting the assumption that exceeding **minimum cutoffs scores** had implications for gains in children's development.
 - **Burchinal (2018)** likewise pointed to psychometric limitations of existing measures.
 - Substantive **variance attributable to raters**, including due to the “within one” inter-rater agreement criterion used by CLASS PreK and ERS.
 - **Low variability on items** makes it difficult to distinguish among classrooms.
 - More attention needed to the **content of ECE activities**, and domain-specific measures when considering growth in particular domains.



Preview: Research and Practice Use

- A major concern is with **high stakes policy uses** of measures.
- But it is important to recognize that the **limitations in evidence** also impact **research** and **practice uses**.
 - Measurement error and limited variation **attenuate associations** with quality scores as outcomes and predictors.
 - Uncertainty of estimation may lead **professional development** to be **less well tailored** to a teacher's current practices and needs.



Preview: Three Key Tensions

- **The value of “humanness” in ratings**
- **The technical view of psychometrics**
- **The money-making potential of measures**



First Tension: An Age Old Challenge

- **Core tension:**
 - **Observations** of early childhood classrooms offer the most **authentic** signal of the quality of teaching practices and adult-child interactions.
 - The **humanness** that makes observations authentic also **challenges inter-rater reliability**, since two people may see the same thing somewhat differently.
- **Related tension:**
 - Trying to **increase inter-rater** reliability by standardizing what raters look for and how they score it can come at the **cost of validity**.
 - Focusing on what is **reliably codable** may **leave out more subjective elements** of classroom practices and interactions that are viewed as most important for children's learning and development.
- **Open question:**
 - To what extent **has the early childhood field fully recognized and grappled with these tensions**, including by drawing on **knowledge accrued in other fields** and by drawing on **modern technology and techniques**?



Second Tension: A Disconnect

- **Core tension:**
 - **Psychometrics** and **measurement theory** have deep and rich traditions that grapple with and help to address core concerns and challenges.
 - **Psychometrics** and **measure validation** are seen as highly technical and specialized, frequently hidden in technical reports and published in specialized journals.
- **Related tension:**
 - Many scientists have **limited training in psychometrics and measurement theory**, especially beyond certain **classical test theory techniques** (percent agreement, Kappa, Cronbach's alpha, basic factor analysis).
 - Justification for measure use can reflect **inertia** and **scale developer summaries**, with local replication of psychometric evidence relatively uncommon.
- **Open question:**
 - To what extent **has the early childhood field fully recognized and grappled with these tensions**, including by reflecting on the ways in which reliance on classical test theory and scale developer evidence has **produced blind spots** and **limited robust continuous measure improvement**.



Third Tension: A Modern Twist

- **Core tension:**
 - **Measures** are an **integral part of operationalizing** conceptual frameworks and logic models, a core aspect of the **scientific enterprise**.
 - **Measures** are **tangible products** that have many uses outside of science, and some markets have led measures to become part of **big business**.
- **Related tension:**
 - **Scientific norms** emphasize **common ownership** of scientific goods, encourage **public benefits**, and continually place claims under **critical scrutiny**.
 - **Commercial norms** emphasize **private ownership** of intellectual property, focus on **private gains**, and **protect assets** from damaging evidence.
- **Open question:**
 - To what extent **has the early childhood field fully recognized and grappled with these tensions**, including by reflecting on the **unintended consequences** of unrecognized and uncontested historical practices and **contemporary movements** toward open science and ownership of personal data?



The Details



Examples of Use and Evidence



Brief Reminder: Public Investments and High Stakes Use of Measures



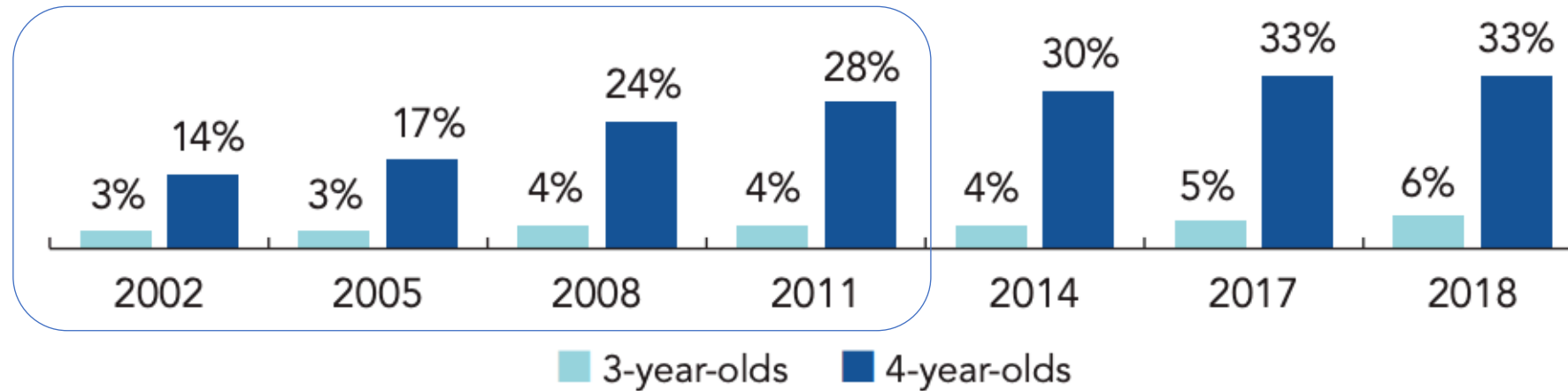
Key Points

- Investments in early care and education grew in recent decades...
- With a focus on the quality of classrooms receiving these investments, particularly in relation to school readiness
- This led to increasing adoption of existing observational measures to assess classroom quality



Expanding Public Investments in Early Care and Education

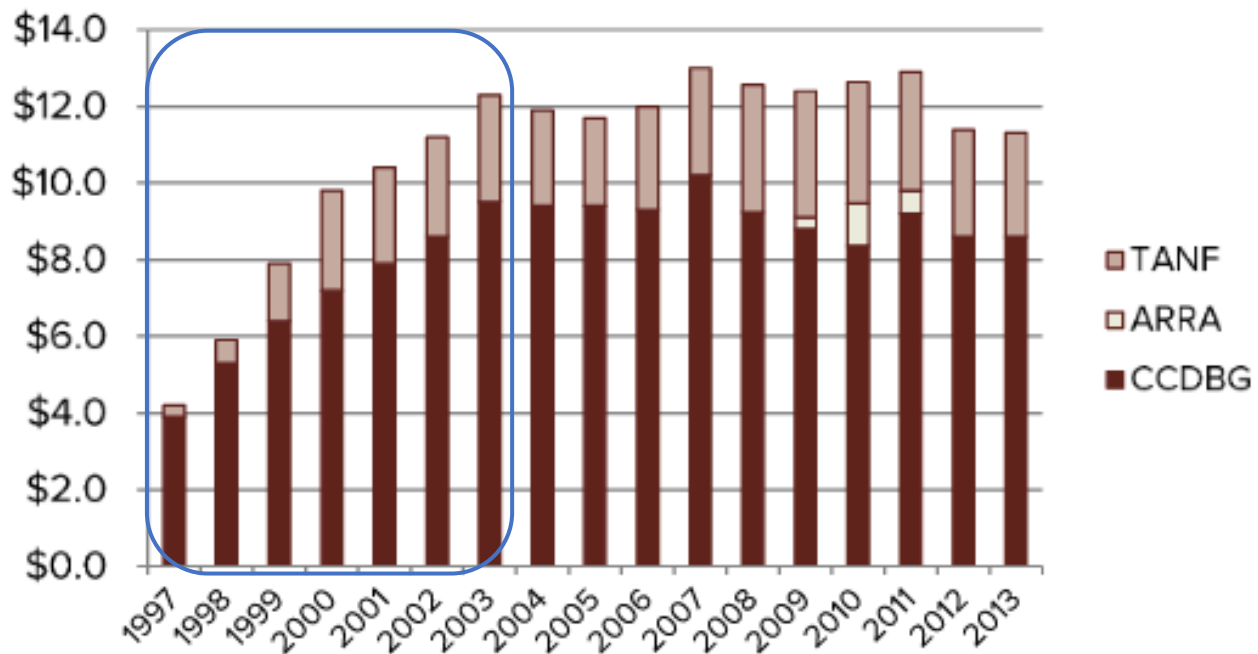
PERCENT OF STATE POPULATION ENROLLED



Percentage of 4 year olds enrolled in state pre-k doubled, 2002 to 2011

Expanding Public Investments in Early Care and Education

Figure 1. Total Combined State and Federal Child Care Spending (in billions), 1997-2013



Source: CLASP calculations based on HHS data

Federal and state child care spending through CCDF/TANF tripled, 1997 to 2003



Expanding Public Investments in Early Care and Education

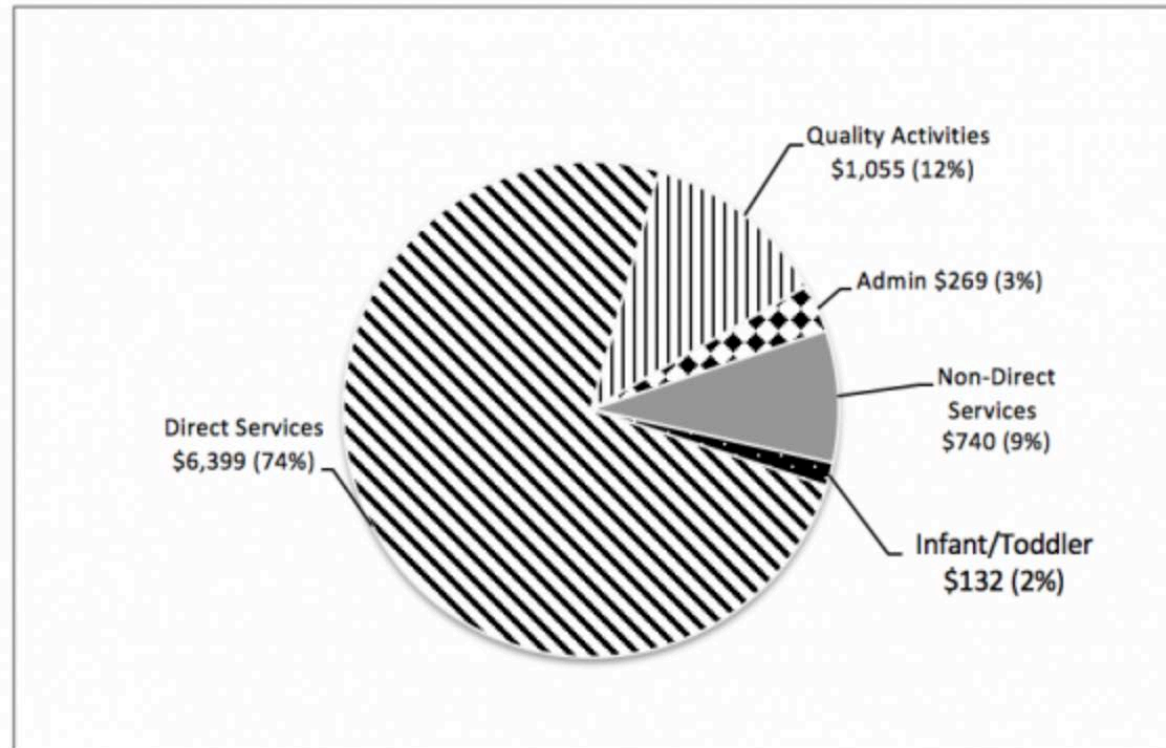


Chart 1 - Total FY 2017 Expenditures by Category (in millions)

CCDF/TANF expenditures in FY2017 totaled \$8.6 billion, with 12% focused on quality activities.



Expanding Public Investments in Early Care and Education

Options for Quality Activities



Policy Focus on Quality Early Care and Education

- **Obama-era** policy initiatives focus on **high-quality** early care and education.
- Typically with at least part of the goal being support for children's **school readiness** and later **school and life success**.
- More recent turn in **Trump era** toward **cost and supply**, including reducing the **burden of regulations**.
- Leading **2020 Democratic Presidential candidates** proposing access to "high-quality" child care programs, with Warren's particularly detailed and modelled on Head Start and military child care.

Obama-Era Focus

FACT SHEET: Invest in US: The White House Summit on Early Childhood Education

- **Providing High-Quality Preschool for Every Child:** The President has proposed a new federal-state partnership to provide all low- and moderate-income four-year old children with high-quality preschool, while also expanding these programs to reach additional children from middle class families and incentivizing full-day kindergarten policies. This investment - financed through a cost-sharing model with states - will help close America's school readiness gap and ensure that children have the chance to enter kindergarten ready for success. Congress provided \$250 million in FY2014 under the Preschool Development Grants program.

Obama-Era Focus

Education Department Announces Next Rounds of Race to the Top, Including Another Key Investment to Expand Access to High-Quality Early Learning Opportunities

APRIL 16, 2013

Contact: Press Office, (202) 401-1576, press@ed.gov

The U.S. Department of Education and U.S. Department of Health and Human Services announced they will invest the majority of the 2013 Race to the Top funds for a second Race to the Top-Early Learning Challenge competition. About \$370 million will be available this year for states to develop new approaches to increase high-quality early learning opportunities and close the school readiness gap. Today's announcement furthers the Administration's work to expand access to high-quality early learning programs for all children, especially those in disadvantaged communities.



Improving Head Start for School Readiness Act of 2007

Public Law 110–134
110th Congress

An Act

To reauthorize the Head Start Act, to improve program quality, to expand access,
and for other purposes.

Dec. 12, 2007

[H.R. 1429]

*Be it enacted by the Senate and House of Representatives of
the United States of America in Congress assembled,*

SECTION 1. SHORT TITLE.

(a) **SHORT TITLE.**—This Act may be cited as the “Improving
Head Start for School Readiness Act of 2007”.

(b) **TABLE OF CONTENTS.**—The table of contents of this Act
is as follows:

Improving Head
Start for School
Readiness Act
of 2007.
42 USC 9801
note.

“(F) include as part of the reviews, a valid and reliable
research-based observational instrument, implemented by
qualified individuals with demonstrated reliability, that
assesses classroom quality, including assessing multiple
dimensions of teacher-child interactions that are linked
to positive child development and later achievement;



Policy Focus on Quality Early Care and Education

- This reflects a **sensible desire** to ensure that public dollars invest in high quality settings.
- But, a desire that is **difficult to put into practice**.
- **Alternative options** for such high stakes use of quality measures.
 - Choose the **relatively best measure available at the time** and use “as is” (even if evidence limited).
 - Choose existing measure but **build in rigorous evidence building** and **potential for modifications** to measure during use.
 - Require an **absolute level of evidence** before use.
- The **first approach was used**, and the ramifications are increasingly recognized.
- The **second and third approaches** are gaining momentum.



ECERS-R and CLASS: Structure and Evidence



Key Points

- At a high level, ECERS-R and CLASS offer counterpoints in relation to the tensions in relation to reliability and validity already noted
- At a deeper level they share several limitations, including in relation to:
 - Rater effects
 - Item variation
 - Standard errors of measurement

ECERS-R and CLASS

- Two most widely used observational measures.
- Similarities and differences:
 - Both have observers visit classrooms for several hours to rate actual classroom activities and interactions.
 - Both produce ratings on a 1 to 7 scale.
 - But, different origins and structures.
 - At first blush, one more “checklist” and other more inferential, but both have subjectivities through human scoring.
 - Both are marketed and sold through companies for their widespread use.



ECERS

- Developed in 1970s from a *checklist* to *help practitioners improve* the quality of their settings.
- Reflects the early childhood education field's concept of *developmentally appropriate practice* (whole child approach, child-initiated activities, teacher facilitation responsive to individual needs)

- ECERS-R: 43 item scores as 400+ indicators.
- New version: ECERS-3.

ECERS-R Item 10: Meals/Snacks						
Inadequate	2	Minimal	4	Good	6	Excellent
1		3		5		7
10. Meals/snacks						
1.1 Meal/snack schedule is inappropriate (Ex. child is made to wait even if hungry).		3.1 Schedule appropriate for children.	→	5.1 Most staff sit with children during meals and group snacks.†	→	7.1 Children help during meals/snacks (Ex. set table, serve themselves, clear table, wipe up spills).
1.2 Food served is of unacceptable nutritional value.*		3.2 Well-balanced meals/snacks.*		→	5.2 Pleasant social atmosphere.	7.2 Child-sized <i>eating</i> utensils used by children to make self-help easier (Ex. children use small pitcher, sturdy serving bowls and spoons).
1.3 Sanitary conditions not usually maintained (Ex. most children and/or adults do not wash hands before handling food; tables not sanitized; toileting/diapering and food preparation areas not separated).		3.3 Sanitary conditions usually maintained †		5.3 Children are encouraged to eat independently (Ex. child-sized eating utensils provided; special spoon or cup for child with disabilities).		7.3 Meals and snacks are times for conversation (Ex. staff encourage children to talk about events of day and talk about things children are interested in; children talk with one another).
1.4 Negative social atmosphere (Ex. staff enforce manners harshly; force child to eat; chaotic atmosphere).		3.4 Nonpunitive atmosphere during meals/snacks.		5.4 Dietary restrictions of families followed. NA permitted.		
1.5 No accommodations made for children's food allergies. NA permitted.		3.5 Allergies posted and food/beverage substitutions made. NA permitted.		3.6 Children with disabilities included at table with peers. NA permitted.		

Sources: Harms, T., Clifford, R.M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale, Revised Edition*. New York, NY: Teachers College Press.



CLASS

- Developed in *1990's/2000's* beginning in a *research* study and later aimed at *professional development and coaching*.
- Reflects *developmental theory and research* and emphasizes teacher-student (adult-child) *interactions* as the primary means of development and learning.
- Observers assimilate what they see and report scores to just a few items.

The manual advises:
“**Because of the highly inferential nature of the CLASS, scores should never be given without referring to the manual.**”
(Pianta, La Paro & Hamre, p. 17, bold in original)



Use in State Quality Rating and Improvement Systems

Among states that use observational measures for their QRIS,
what percentage use ERS and CLASS for some purpose (rating, quality improvement)
perhaps along with another tool?

ERS = Suite of measures for preschools (ECERS-R), infant/toddler centers (ITERS-R) and homes (FCCERS-R).

2017:

~ 75% ERS

~55% CLASS

Most recently, **over half use CLASS**
and about **three-quarters use ERS**.

2010

~ 88% ERS

~ 7% CLASS



Growth in use of CLASS evident
when compared to 2010.



Example of Cutoff Scores: Illinois' QRIS Learning Environment

Program demonstrates high quality of classroom environment

LICENSED CHILD CARE CENTER	PRESCHOOL	INFANT/TODDLER CARE AND EDUCATION	START
<p>ERS¹ ave At least 4 below 4.0 assessm assessor <i>OR</i> _____ CLASS² E Classroom average s no classroom below 4.0 as verified by on-site assessment by state-approved assessor³ <i>OR</i> _____ Accredited sites: Evidence from state-approved accrediting body</p>	<p>with no classroom verified by on-site a by state-approved</p>	<p>score: classroom y on-site Start</p>	<p>pport and on: average no verified by Head sor performance oliance:</p>



1304.21(a)(1) 1304.21(a)(3)

Example of Cutoff Scores: Head Start Designation Renewal System

What do the Head Start CLASS® review scores mean?

^ Q: What CLASS® scores cause a grantee to be required to compete?

A: There are two circumstances under which a grantee is required to compete as the result of low CLASS® scores. First, grantees with average CLASS® scores below the established minimum on any of the three CLASS® domains will be required to compete. These thresholds have been established as a score of 4 for the domain of Emotional Support, 3 for the domain of Classroom Organization, and 2 for the domain of Instructional Support. Second, each year the 10 percent of grantees reviewed that receive the lowest average scores in each domain are required to compete.

If a program scores in the bottom 10 percent of all Head Start programs, this means that the vast majority of Head Start programs were assessed at higher levels. However, if the lowest 10 percent in any of the three CLASS® domains should include grantees with a score of 6 or 7, those grantees would not be required to compete, even if this means that fewer than 10 percent would be required to compete based on that domain.

^ Q: What was the threshold for the lowest 10 percent of CLASS® scores in 2017 by domain?

A: Grantees that had a review conducted in 2017 and that had scores less than or equal to the numbers below are in the lowest 10 percent in each respective CLASS® domain:

- Emotional Support - 5.7024
- Classroom Organization - 5.3264
- Instructional Support - 2.3095

Highlights: Evidence for High Stakes Use



ECERS-R and CLASS

- What is their **evidence**?
 - Do scores on each measure **predict large school readiness gains**?
 - Do measures **sharply define constructs** aligned with readiness gains?
 - Are measures constructed for **maximal precision** (high signal vs. noise)?
- Important and instructive to consider some of the details **at the time they were adopted** (to reflect on the extent to which the evidence at the time did not support their high stakes use).
- **And how certainty in their limitations have grown over time.**



First Limitation of Evidence:
Do ECERS-R and CLASS predict large
school readiness gains?



Summary of Evidence

- Evidence around the time high stakes measures were being adopted showed often **nonsignificant and small** associations between ECERS-R and CLASS PreK scores and children's developmental outcomes.
 - But field historically had tended to focus on **significance rather than size** of associations.
 - And, field had tended to focus on the **few effects that were significant** (“quality can associate with...”) rather than the **more numerous effects that were nonsignificant**.
- That evidence has been further **accumulating over time** as scholars have explored various reasons for these limitations (e.g., threshold effects) and have used modern meta-analyses to emphasize fuller body of evidence and replication across samples).



Example of a Readiness Gap

- **Peabody Picture Vocabulary Test (PPVT)**
- **Standard Deviation:** Standard deviation is 15 points in norming sample
- We find it is ~16 in national Head Start “FACES” samples.
- **Real World Benchmark:** The gap between lower and higher income kindergartners in one representative survey of children from large cities (Fragile Families) was approximately 15 points.
- *Suggests that the PPVT can pick up meaningfully large school readiness gaps.*
- *And that associations with quality measures would have to be sizable to “close the gap” as in the goals of policy where we see high stakes use of ECERS-R and CLASS PreK.*
- *For instance, standardized coefficients (approximate effect sizes) would need to approach 1 for a one standard deviation increase in quality to close the gap.*



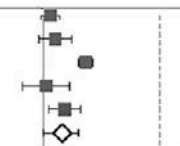
Recent Meta-Analyses of Published Studies

Simple Correlations

Associations often nonsignificant and meta-effects below .10 in size.

CLASS PreK Instructional Support & PPVT

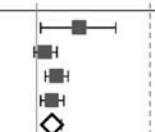
Outcome Variable	Source	Correlation (95% CI)	Sample Size
PPVT – Vocabulary	Aikens 2012[27]	0.03 (-0.01 to 0.07)	1936
	Burchinal 2014[34]	0.05 (-0.02 to 0.12)	822
	Dotterer 2012[39]	0.18 (0.15 to 0.21)	3584
	Weiland 2013[57]	0.01 (-0.09 to 0.11)	414
	West 2010[58]	0.09 (0.02 to 0.16)	684
		0.08 (0.00 to 0.15), $I^2=0.0\%$	



Perlman et al. 2016

ECERS/ECERS-R Language Reasoning & PPVT

Outcome Variable	Source	Correlation (95% CI)	Sample Size
PPVT - Vocabulary	Burchinal 2011[15] - CQO	0.19 (0.02 to 0.35)	140
	Burchinal 2011[15] - FACES 1997	0.04 (-0.01 to 0.09)	1493
	Burchinal 2011[15] - FACES 2000	0.09 (0.04 to 0.14)	1739
	Burchinal 2011[15] - NCEDL	0.07 (0.02 to 0.12)	1465
		0.07 (0.04 to 0.11), $I^2=8.7\%$	



Brunsek et al. 2017

Similar results are seen for other subscales and outcomes.



Example: Head Start and CLASS

We examined regression-adjusted standardized associations of each CLASS subscale with PPVT in the Head Start nationally representative FACES 2009 and 2014 studies.

Associations were nonsignificant and small
(.03 to .04 for Instructional Support subscale)

FACES 2014	
	PPVT
Total Raw Score	-0.01 (0.05)
Raw Average (3d Standard)	
Emotional Support (Items 1,2,3,4)	0.00 (0.08)
Classroom Organization (Items 5,6,7)	-0.04 (0.10)
Instructional Support (Items 8,9,10)	0.03 (0.04)

FACES 2009	
CLASS Scores	PPVT
Total Raw Score	0.01 (0.02)
Raw Average (3d Standard)	
Emotional Support (Items 1,2,3,4)	-0.04 (0.04)
Classroom Organization (Items 5,6,7)	0.03 (0.04)
Instructional Support (Items 8,9,10)	0.04 (0.03)



Example: Head Start and CLASS

Other outcomes: Also typically nonsignificant and consistently small associations

Range of outcomes

- Woodcock Johnson
 - Letter Word
 - Spelling
 - Applied Problems
 - Pencil Tapping Inhibitory Control
 - Leiter Attention/Social
 - Teacher-Reported (and Parent-Reported)
 - Social Skills
 - Behavior Problems
 - Spanish TVIP/ROWPVT
 - Woodcock-Muñoz
 - Letter Word
 - Spelling
 - Applied Problems
- 95% of associations were non-significant.
 - Level would expect by chance with Type I error.
 - The potential these associations reflect chance was reinforced since which associations were significant did not replicate between 2009 and 2014.
 - Significant associations were small.
 - $|.14|$ and below.
 - Not always in conceptually expected directions.



Do ECERS-R and CLASS sharply
measure constructs aligned with
readiness gains?



Importance of Dimensions of Quality

- Ideally, measures would be created specifically for **aspects of quality aligned with policy goals**.
 - For the school readiness goals, would want **content-focused** aspects of quality aligned with **particular readiness domains**.
- If measures were designed for other purposes, they should still have **clear definitions** of the **aspects of quality** measured and **empirical evidence** that **items well cover** those dimensions.
 - When the dimensions/domains are written into high stakes policy, they become a **focal point** for teachers and programs.
 - Using cutoffs connected to the dimensions **implies the dimensions are meaningful and relevant** to the policy goals.



ECERS-R Dimensions: One, Seven, or Two (Three)?

- The ERS presume a quality program supports **three basic needs** (health/safety, positive relationships, opportunities for learning from experience) and “no one is more or less important than the others.”
 - The ECERS-R scale developers sometimes describe the measure as capturing a **single global aspect of quality**.
 - But items are organized into **seven subscales**, some of which on the surface align with particular aspects of quality (personal care, interaction, activities).
 - **Some QRIS**, like Illinois, rely on either the total or subscale scores.
- On the other hand, **factor analyses** have identified **2-3 dimensions**.
 - These dimensions are **sometimes used in QRIS**.



New Version: ECERS-3

- ECERS-3 has 6 subscales.
- But a recent publication identified 4 factors.
- Std. associations with children’s outcomes were **generally nonsignificant and consistently small.**
 - **83% nonsignificant**
 - Significant associations **.08 or smaller.**

Table 8

Standardized parameter estimates (standard errors) for interaction of time by ECERS-3 scores as predictors of social–emotional and academic skills.

Dependent variable	ECERS-3				
	Total Score	Learning Opportunities	Gross Motor	Teacher Interactions	Math Activities
DECA Total Protective T-score (<i>n</i> = 533)	0.01 (0.04)	0.02 (0.04)	0.02 (0.04)	0.00 (0.04)	0.08 (0.04)*
DECA Behavioral Concerns T-score (<i>n</i> = 533)	–0.03 (0.03)	–0.03 (0.04)	–0.05 (0.04)	–0.03 (0.03)	–0.02 (0.03)
HTKS Total Score (<i>n</i> = 572)	0.06 (0.03)*	0.08 (0.03)**	–0.03 (0.04)	0.06 (0.03)*	0.03 (0.04)
WJ IV Picture Vocabulary W Score (<i>n</i> = 575)	0.00 (0.03)	–0.02 (0.03)	0.00 (0.03)	0.01 (0.03)	0.00 (0.03)
WJ IV Letter-Word W Score (<i>n</i> = 575)	0.05 (0.03)†	0.05 (0.03)†	0.05 (0.03)†	0.01 (0.03)	0.04 (0.02)†
WJ IV Applied Problems W Score (<i>n</i> = 575)	0.03 (0.03)	0.08 (0.03)*	–0.01 (0.03)	0.01 (0.03)	–0.02 (0.03)

Notes: Significant associations appear in bold. For all outcomes other than HTKS, each cell represents a separate 3-level HLM in which time (pre- vs. posttest) is nested within child, which is nested within classroom, and the parameter estimates presented are for the interaction of time by ECERS-3. For HTKS, each cell is a 2-level HLM, in which child is nested in classroom, and pretest score is controlled. The parameter estimates presented for HTKS are for the effect of ECERS-3. For all models, the parameter estimates have been standardized so that they represent the amount of growth on the dependent variable, in standard deviations, associated with a one standard deviation change on the CLASS Pre-K.

* *p* < .05.

** *p* < .01.

† *p* < .10.



CLASS PreK Dimensions: Three Domains or a “Bi-Factor”?

- CLASS PreK manual produces scores in **three broad domains**:
 - Emotional Support
 - Classroom Organization
 - Instructional Support
- Due to limited evidence for these three domains, CLASS developers published a **“bi-factor” structure** for the CLASS PreK (Hamre et al., 2014) that differs from the subscales written into policy.
 - **One general** dimension (Responsive teaching)
 - **Two specific** dimensions
 - Proactive management and routines
 - Cognitive facilitation



CLASS PreK Dimensions: Three Domains or a “Bi-Factor”?

- We replicated this bi-factor structure in the Head Start FACES 2009 and 2014 samples.
 - Like the CLASS developers, however, we had **problems with convergence.**
- **Alternative traditional structures** fit as well or better.
 - 2-dimensional: combined ES and CO items into one factor.
 - 3-dimensional reconfigured ES and CO items.



Associations with Child Outcomes Still Generally Nonsignificant and Consistently Small

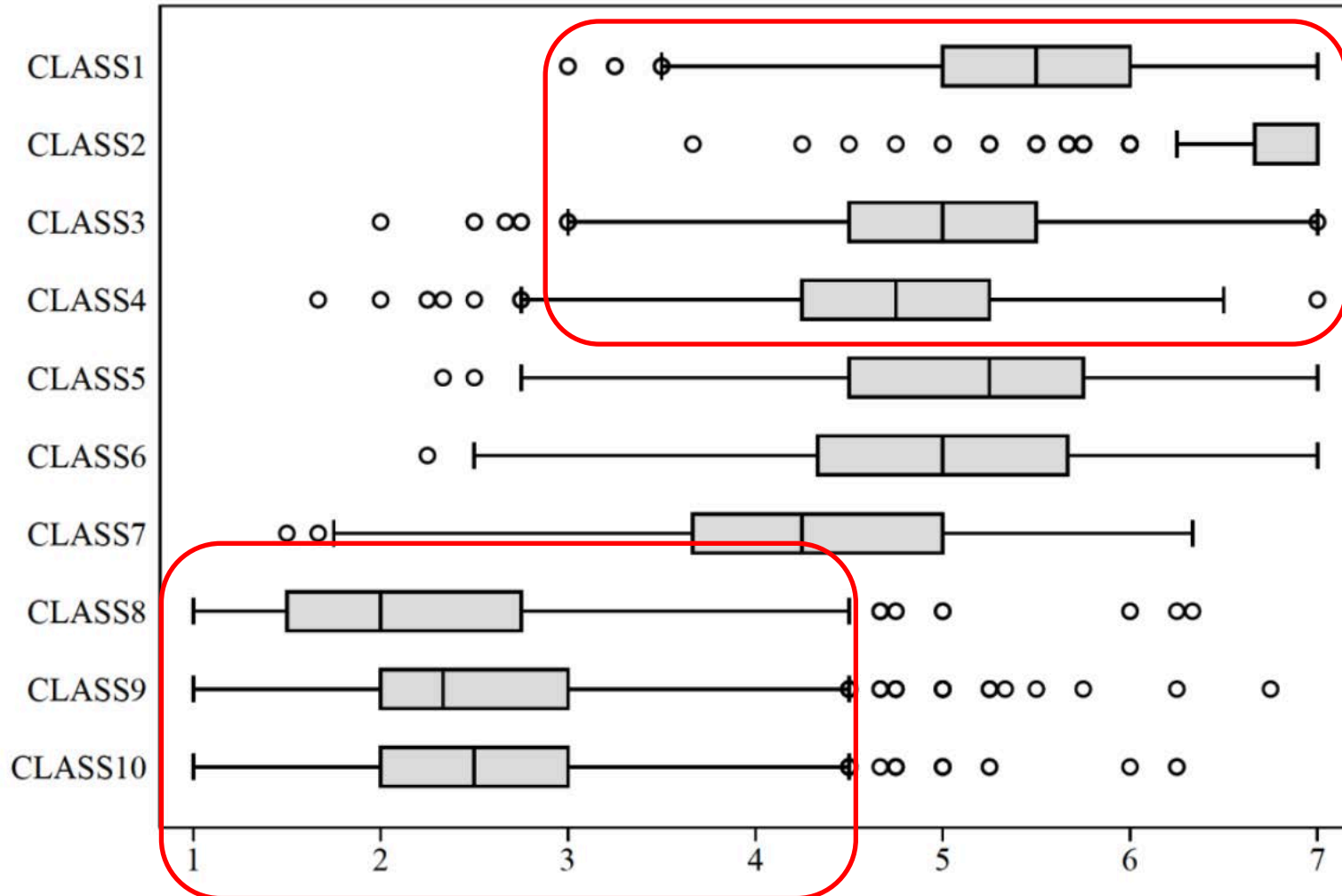
	FACES 2014
	PPVT
Total Raw Score	-0.01 (0.05)
Raw Average (3d Standard)	
Emotional Support (Items 1,2,3,4)	0.00 (0.08)
Classroom Organization (Items 5,6,7)	-0.04 (0.10)
Instructional Support (Items 8,9,10)	0.03 (0.04)
Raw Average (3d Alternative) ^a	
Climate & Management (Items 1,2,5,6)	-0.09 (0.07)
Sensitivity & Regard (Items 3,4,7)	0.06 (0.07)
Raw Average (2d Alternative) ^a	
Combine ES & CO (Items 1-7)	-0.03 (0.05)
Confirmatory Bifactor	
General	-0.01 (0.05)
Proactive Mgt. & Routines (Items 1-7)	-0.02 (0.04)
Cognitive Facilitation (Items 8-10)	0.02 (0.04)

	FACES 2009
CLASS Scores	PPVT
Total Raw Score	0.01 (0.02)
Raw Average (3d Standard)	
Emotional Support (Items 1,2,3,4)	-0.04 (0.04)
Classroom Organization (Items 5,6,7)	0.03 (0.04)
Instructional Support (Items 8,9,10)	0.04 (0.03)
Raw Average (3d Alternative) ^a	
Climate & Management (Items 1,2,5,6)	0.01 (0.03)
Sensitivity & Regard (Items 3,4,7)	-0.03 (0.03)
Raw Average (2d Alternative) ^a	
Combine ES & CO (Items 1-7)	-0.02 (0.03)
Confirmatory Bifactor	
General	0.00 (0.02)
Proactive Mgt. & Routines (Items 1-7)	-0.02 (0.02)
Cognitive Facilitation (Items 8-10)	0.03 (0.02)



Example: Distribution of Scores in Head Start

FACES 2014



Item skewness may contribute to convergence issues.

Emotional support items concentrated in the higher quality regions.

Instructional support items concentrated in lower quality regions.

This limited item variation is also especially problematic for the “within one” agreement used for certification, which I’ll discuss in a moment.

Will see that within-one range overlaps observed score range for majority of cases.



Are ECERS-R and CLASS
constructed for maximal
precision (high signal vs. noise)?



Scoring Strategies May Produce Noise

- The structures of ECERS-R and CLASS are quite different, but each may increase noise.
 - ECERS-R **checklist origin** of 400+ indicators, but used “**stop scoring**” which reduces burden so not all need to be rated.
 - CLASS a **highly inferential** approach, where coders assimilated all they’ve seen in their heads, rather than explicitly scoring and numerically summarizing indicators and markers.
- **At first blush**, these structures seem to reflect **the tradeoffs of “humanness”** of ratings discussed earlier (reliability vs. validity).
- But, the **wording of ECERS-R** indicators often **require interpretation** (hence a 100 plus page **All About the ECERS-R** step-by-step guide to scoring).

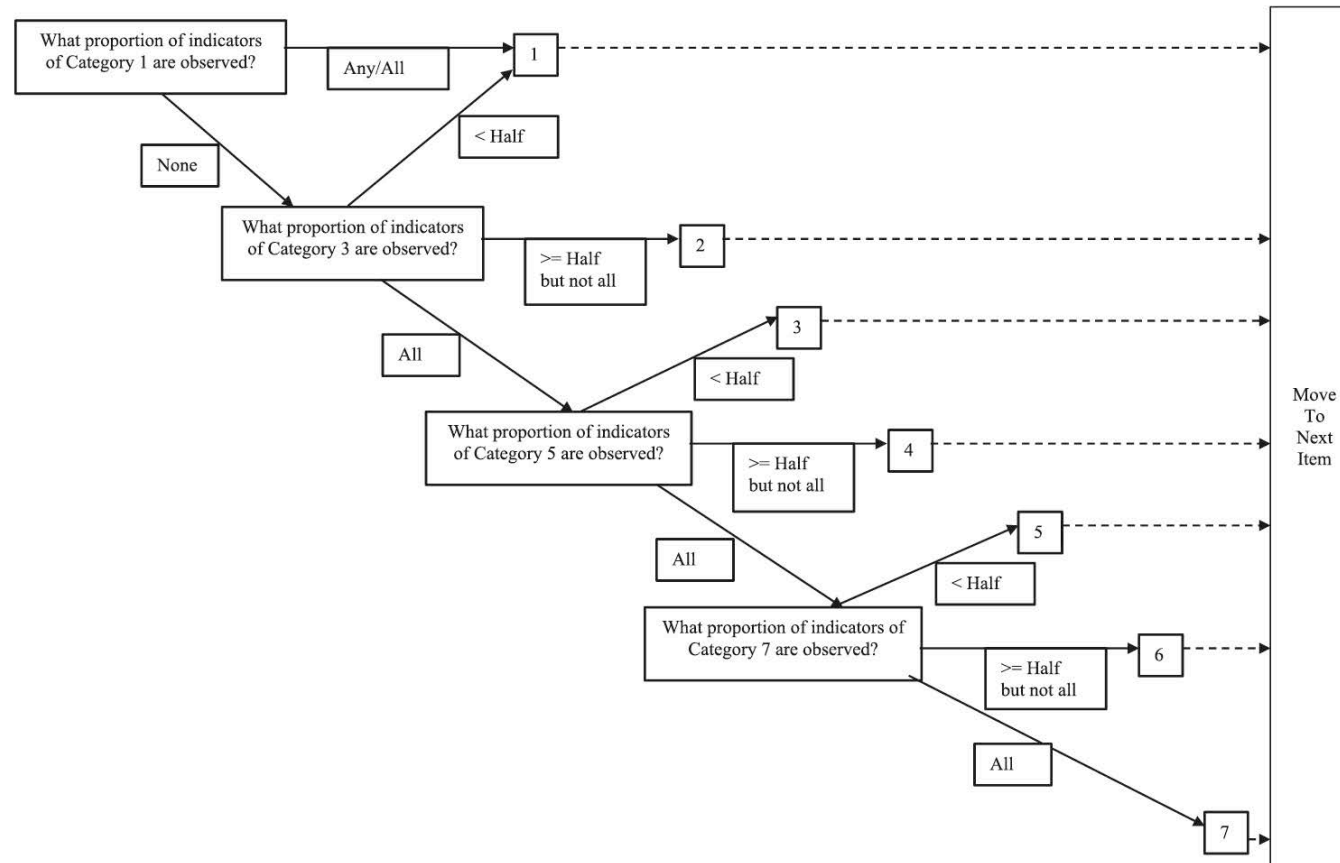


ECERS Standard “Stop Scoring”



ECERS-R Standard “Stop Scoring”

- **Conditions** in the indicators of **lower scores** must be met before indicators of higher scores are evaluated.
- **Rules** differ for **even and odd** scores.



Fujimoto et al.
2018

FIGURE 1. Visual representation of the Early Childhood Environment Rating Scale, Revised (ECERS-R) stop-scoring guidelines.

Note. Indicators of Category 1 are negatively oriented. Indicators of Categories 3, 5, and 7 are positively oriented.

Example of Possible Issues with Stop Scoring

ECERS-R Item 10: Meals/Snacks

Indicators placed based on scale developers' philosophical and practice lens regarding quality.

Different aspects of quality are mixed together across indicators.

A higher score may reflect more of some but not all aspects of quality.

Hard to pull out aspects of quality aligned with certain readiness goals.

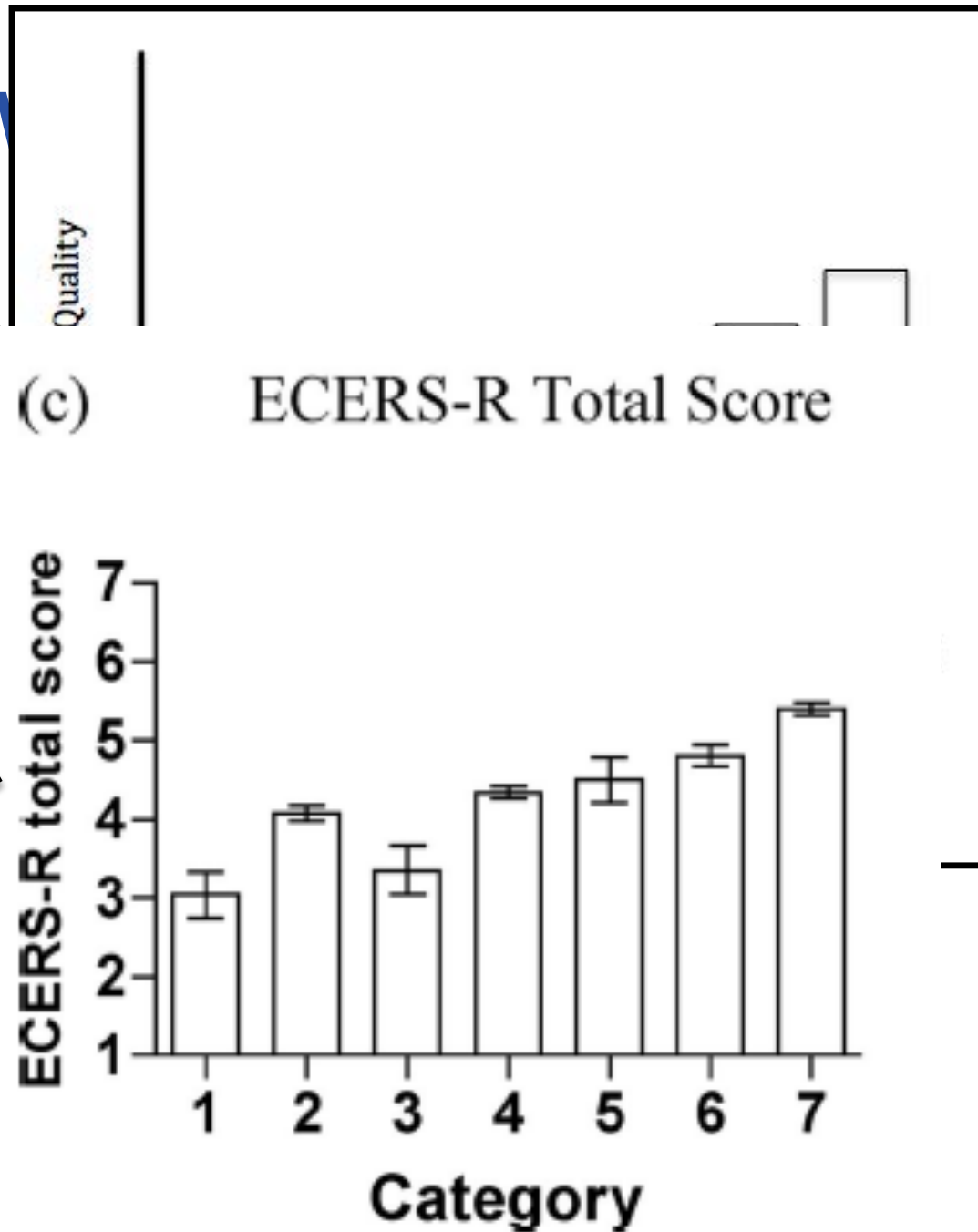
Inadequate 1	2	Minimal 3	4	Good 5	6	Excellent 7	
10. Meals/snacks							
1.1 Meal/snack schedule is inappropriate (Ex. child is made to wait even if hungry).		3.1 Schedule appropriate for children.	→	5.1 Most staff sit with children during meals and group snacks.‡	→	7.1 Children help during meals/snacks (Ex. set table, serve themselves, clear table, wipe up spills).	
1.2 Food served is of unacceptable nutritional value.*		3.2 Well-balanced meals/snacks.*		→	5.2 Pleasant social atmosphere.	7.2 Child-sized <i>servicing</i> utensils used by children to make self-help easier (Ex. children use small pitcher, sturdy serving bowls and spoons).	
1.3 Sanitary conditions not usually maintained (Ex. most children and/or adults do not wash hands before handling food; tables not sanitized; toileting/diapering and food preparation areas not separated).		3.3 Sanitary conditions usually maintained.†		5.3 Children are encouraged to eat independently (Ex. child-sized <i>eating</i> utensils provided; special spoon or cup for child with disabilities).		→	7.3 Meals and snacks are times for conversation (Ex. staff encourage children to talk about events of day and talk about things children are interested in; children talk with one another).
1.4 Negative social atmosphere (Ex. staff enforce manners harshly; force child to eat; chaotic atmosphere).		3.4 Nonpunitive atmosphere during meals/snacks.		5.4 Dietary restrictions of families followed. <i>NA permitted.</i>			
1.5 No accommodations made for children's food allergies. <i>NA permitted.</i>		3.5 Allergies posted and food/beverage substitutions made. <i>NA permitted.</i>					
		3.6 Children with disabilities included at table with peers. <i>NA permitted.</i>					

Source: Harms, T., Clifford, R.M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale, Revised Edition*. New York, NY: Teachers College Press.

What W

- If higher score
quality score;
higher category

- Alternatively,
dips in average
scores.



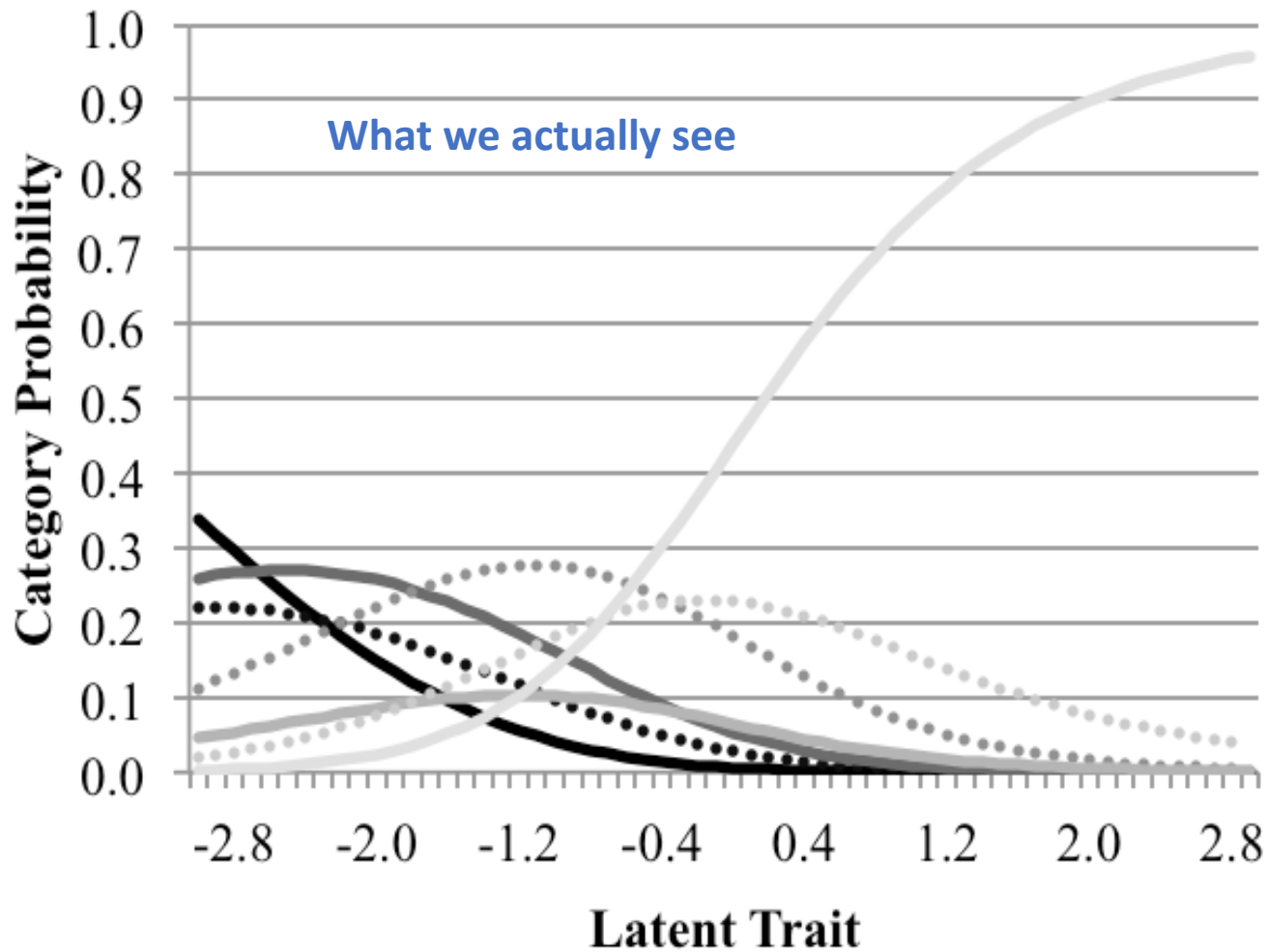
IRT Models Confirm Lack of Order

- We analyzed data from eight studies
 - 14 waves of data.
 - 4,000 classrooms.
- We used multiple kinds of models with various assumptions
 - Nominal response model.
 - Generalized partial credit model.
 - Partial credit model.
 - Within-category average scores.
 - Point-biserial correlations.
- We identified problems with the assumptions for all items.
 - All 36 items had categories that did not follow an ordinal progression with respect to quality.
 - One-fifth had categories fully out of order.
 - The category problems accumulated to the scale score level.
 - The results caution against the use of the standard raw scoring.

Fujimoto et al.
2018



A
(Anc



Is
familiar)

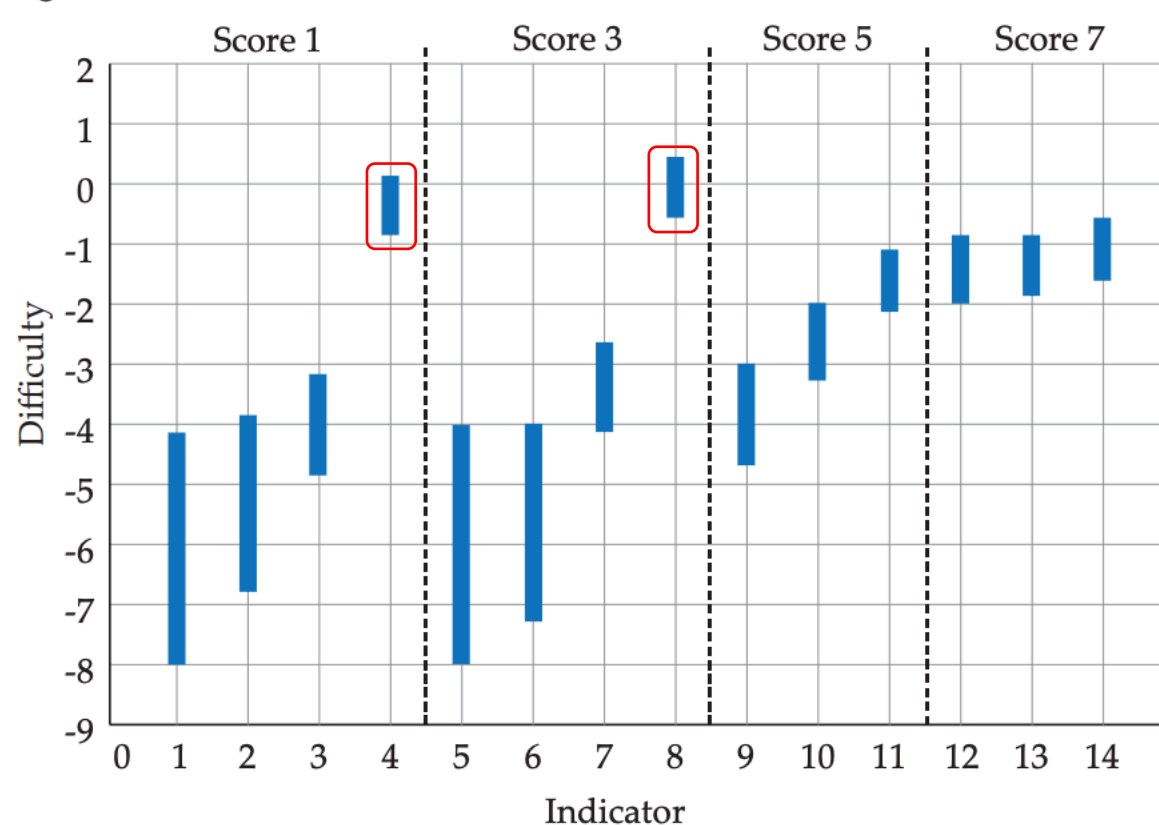
- 1
- 2
- 3
- 4
- 5
- 6
- 7



How the Problems Happen

When we were able to analyze the underlying indicators, we confirmed that some placed at lower scores were in fact “harder” (higher on the latent construct) than some places at higher scores.

Figure 2: ECERS-R 10: Meals/Snacks²



Indicator Key

1. Acceptable nutritional value
2. Appropriate meal schedule
3. Positive atmosphere
4. Sanitary condition
5. Well-balanced meals
6. Schedule appropriate
7. Nonpunitive atmosphere
8. Sanitary conditions usually maintained
9. Eat independently
10. Pleasant atmosphere
11. Staff sits with children
12. Conversation
13. Child-sized utensils
14. Children help

Sanitary conditions high hurdle to get over in order to “get credit” for social and conversational aspects of meals.

Gordon et al. 2015



Implications

- These results **caution against the way item averages are used**, including in high stakes cutoffs.
- ***“when category distinctions fail to discriminate, a researcher would not want to use a scoring strategy that aggregates raw integer item scores”*** Preston and Reise (2015, p. 392)



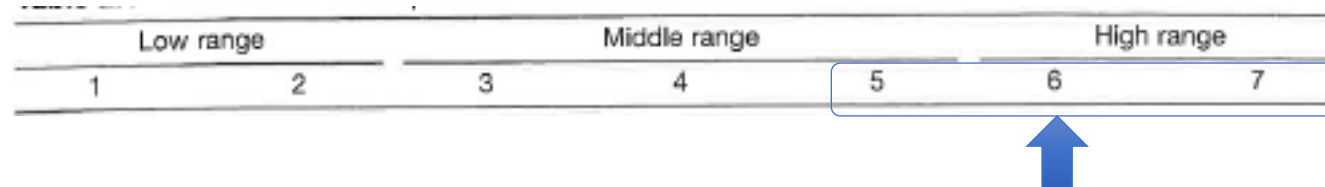
CLASS Inferential Scoring



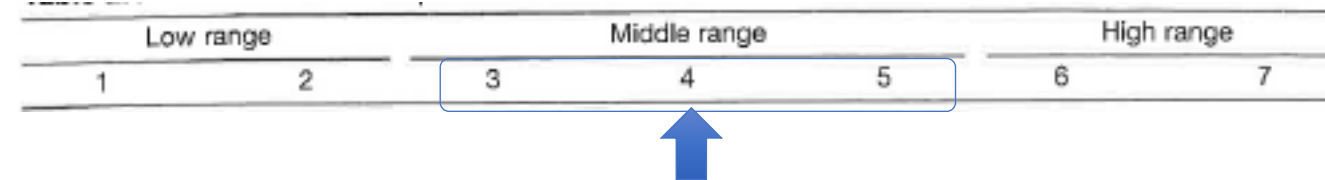
CLASS Inter-Rater Reliability: Is “Within One” Good Enough?

For certification, the CLASS (like the ECERS-R) assesses agreement “within one” on the 7-point scale.

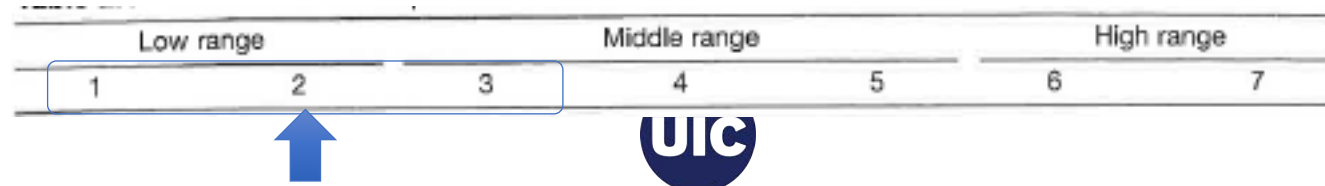
A score of 5, 6 or 7 is considered in agreement with a master score of 6.



A score of 3, 4 or 5 is considered in agreement with a master score of 4.



A score of 1, 2 or 3 is considered in agreement with a master score of 2.



As already noted, given the skewness of the CLASS item scores, this within one agreement covers similar ground as the typical item ranges.

Within one historical use reflected fact that exact agreement is difficult to achieve for rater-mediated assessments, especially on a highly inferential system

Again, CLASS observers have to assimilate considerable information, and use their own judgment in relating what they see to the manual narratives for each level of each item.

Table 2.1. Dimension descriptions for the CLASS

Low range		Middle range	
1	2	3	4
The low-range description fits the classroom and/or teacher very well. All, or almost all, relevant indicators in the low range are	The low-range description mostly fits the classroom and/or teacher, but there are one or two indicators that are in the	The middle-range description mostly fits the classroom and/or teacher, but there are one or two indicators in the low	The middle-range description fits the classroom and/or teacher very well. All, or almost all, relevant indicators in the middle

Low Quality of Feedback (1, 2)

The teacher rarely provides scaffolding to students but rather dismisses responses or actions as incorrect or ignores problems in understanding. In scaffolding, a teacher acknowledges where a student is starting and provides the necessary level of help to allow the student to succeed or complete a task. The teacher in the low range of this dimension tends to move quickly during lessons and fails to use hints or assistance when students do not understand something or give an incorrect answer. For example, the teacher may ask a question to a large group of students; when most of the students respond out loud with the incorrect answer, she simply provides the correct answer and moves on. As another example, when asked whether a character in a story is a mom or a teacher, a student incorrectly responds "a mom." Rather than asking the student how he might know whether the character is a mom or a teacher or giving hints, the teacher simply says, "No, it's a teacher." Alternately, the teacher may completely ignore this response from the student and ask another student for her response.

The teacher gives only perfunctory feedback to students. The teacher may not interact with students in a way that allows him or her to provide feedback. For example, the teacher may spend all of an allotted amount of time reading a book and not ask any questions, thus providing no opportunities for feedback. Alternately, he or she may give a lot of feedback but focus entirely on whether an answer is correct, saying "yes" or "no" or "that's not right," and moving on. Teachers at the low end of the Quality of Feedback dimension also may appear to answer all of their own questions, thus not allowing the provision of feedback on students' thoughts and ideas. For example, the teacher may say, "Well, what do you see in this picture? There are some people and some animals in the red barn." The teacher does not engage in a back-and-forth exchange with students intended to help them understand or to elicit a higher level of performance.

Quality of Feedback⁹

Assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation

Low (1,2) Mid (3,4,5) High (6,7)

Scaffolding

- Hints
- Assistance

Feedback loops

- Back-and-forth exchanges
- Persistence by teacher
- Follow-up questions

The teacher rarely provides scaffolding to students but rather dismisses responses or actions as incorrect or ignores problems in understanding.

The teacher occasionally provides scaffolding to students but at other times simply dismisses responses as incorrect or ignores problems in students' understanding.

The teacher often scaffolds for students who are having a hard time understanding a concept, answering a question, or completing an activity.

The teacher gives only perfunctory feedback to students.

There are occasional feedback loops—back-and-forth exchanges—between the teacher and students; other times, however, feedback is more perfunctory.

There are frequent feedback loops—back-and-forth exchanges—between the teacher and students.

Thought processes

...s to explain thinking
...ponses and actions

The teacher rarely queries the students or prompts students to explain their thinking and rationale for responses and actions.

The teacher occasionally queries the students or prompts students to explain their thinking and rationale for responses and actions.

The teacher often queries the students or prompts students to explain their thinking and rationale for responses and actions.

Information

...back

Encouragement and affirmation

...nt
...istence

The teacher rarely provides additional information to expand on the students' understanding or actions.

The teacher occasionally provides additional information to expand on the students' understanding or actions.

The teacher often provides additional information to expand on students' understanding or actions.

The teacher rarely offers encouragement of students' efforts that increases students' involvement and persistence.

The teacher occasionally offers encouragement of students' efforts that increases students' involvement and persistence.

The teacher often offers encouragement of students' efforts that increases students' involvement and persistence.

...feedback is generally observed in response to a student's or students' answer to a student progresses on his or her work or involvement in an activity, whereas
...or activities.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment Scoring System – PreK Manual*. Baltimore, MD: Brookes Publishing.

QUALITY OF FEEDBACK

QUALITY OF FEEDBACK



Challenge of Rater Variance

- **As the Head Start DRS was first being rolled out in 2008 to 2009, CLASS developers (Cash, Hamre, Pianta, & Myers, 2012) reported based on training over 2,000 Head Start staff:**
 - Exact agreement was low.
 - **41% overall exact agreement** with master score.



We similarly see low exact agreement in a team of 14 Raters coding 425 Video Cycles

All trained by a Teachstone certified master trainer.
All passed Teachstone certification.

Coded video for 425 cycles collected with 2 cameras to approximate live observation (one panoramic, one close up).

Inter-Rater Percent Agreement and Kappa Statistics for Each CLASS PreK Item

	All Ratings				I
	Percent Agreement		Fleiss' Kappa ¹		
	Exact	Within One ²	Exact	Within One	
Positive Climate	37.15	79.61	.13	.33	
Negative Climate	65.20	65.20	.28	.28	
Teacher Sensitivity	29.90	73.15	.08	.28	
Regard for Student Perspectives	25.87	73.04	.04	.31	
Behavior Management	34.91	76.65	.12	.30	
Productivity	26.61	68.51	.02	.10	
Instructional Learning Formats	29.25	73.52	.06	.26	
Concept Development	43.94	91.44	.13	.57	
Quality of Feedback	32.59	81.99	.04	.37	
Language Modeling	33.19	77.51	.03	.19	

Median **within one** agreement was **75%**.

Median **exact agreement** was **33%**.

Median **weighted Kappa** was **.29**.

Median **Kappa** was **.07**.

Note. $n = 614$ ratings total

¹ Fleiss' Kappa is the most appropriate kappa coefficient for data with more than two raters.

² "Within one" columns show the agreement coefficients when the raters are allowed to have disagreement within one.



Generalizability Study Results

G-study Results

Sources of variance	Emotional Support	Classroom Organization	Instructional Support
Shorthand	% Total	% Total	% Total
Segment	15	18	26
Rater	14	19	29
Item	39	25	10
Item x Segment	4	5	4
Item x Rater : Segment (Residual)	28	33	31

Note. s = segment. r = rater. i = item. : denotes nesting of one facet in another. The G-theory analysis was based on segments rated by at least 2 raters. $n = 248$ segments. $n = 4$ Emotional Support, 3 Classroom Organization, and 3 Instructing Support items.

- A **g-study** analysis of these data identified substantial variance due to rater.
- **With pure item variance** included in calculations, these **rater-related components account for about 40-60% of the variance.**

Generalizability Study Results

G-study Results

Sources of variance	Emotional Support	Classroom Organization	Instructional Support
Shorthand	% Total	% Total	% Total
Segment	25	24	29
Rater	22	26	33
Item	-	-	-
Item x Segment	7	7	4
Item x Rater : Segment (Residual)	46	44	35

- **Without pure item variance, they account for 68-70% of the variance.**

Note. s = segment. r = rater. i = item. : denotes nesting of one facet in another. The G-theory analysis was based on segments rated by at least 2 raters. $n = 248$ segments. $n = 4$ Emotional Support, 3 Classroom Organization, and 3 Instructing Support items.

Nesting of Segments in Classrooms

- One additional **complication** made salient by a g-study design is the way in which observation **cycles are nested within classrooms**.
- Observational measures, and their **high stakes use**, have not well grappled with the fact that **teachers/classrooms' scores** reflect in part:
 - the **materials/resources available** to them.
 - the **unique strengths and needs of attending children**.



Summary: Limits of Adopting Existing Measures for High Stakes Use

- Limitations in evidence, *including for high stakes policy decisions*.
 - Often nonsignificant and consistently small associations with children's outcomes.
 - Lack of clear domain structures.
 - Problems with scoring.
 - Limited item variation.
 - Low exact rater agreement.
- These results do not necessarily mean that the theoretical and practice models underlying these scales are wrong.
- But, as *operationalized*, the measures have important limitations.



New Strategies for Improving Observational Measures of Classroom Quality



Resolving Tensions

- **Humanness of Ratings:** Use a broader suite of measure development approaches, including those applied to “rater-mediated” measures.
- **Technical View of Psychometrics:** Leverage research approaches to research-practice-policy implementation science.
- **Measures as Products:** Apply modern standards of transparency, replication, and reproducibility.



These are interrelated!

Shifting Thinking: Modern Standards



Modern Standards

- Consistent with the latest *Standards for Educational and Psychological Testing* consider:
 - the ***intents*** of each research, practice and policy use,
 - weigh the ***full body*** of reliability and validity evidence against each use,
 - build in ***continuous and local validation*** of measures selected for these uses,
 - ***allow for the refinement*** of measures over place and time.



In other words

- A measure is **not statically “reliable and valid.”**
 - Such **“sound bite” language** in rules and regs may have the unintended consequence of viewing them as such.
 - And seeing the measures as a **“product” stamped “reliable and valid.”**
- Instead, the **evidence should be fully evaluated and regularly revisited (including locally)** for each use.



For instance

- If it is desirable to distinguish classrooms that fall above and below **specific cutpoints**, as in current policy uses, then measures with **very high information (and low error) at those cutpoints are needed**.
- If the policy goal is to **improve children's school readiness**, then agreement is needed on definitions of readiness and the **aspects of quality** that support them, and measures are needed that are designed and evaluated to assess those aspects of quality.

And, Continuous and Local Validation Means

- The measures go through continuous improvement and local validation.
- This approach **can benefit from viewing** measures as:
 - **Not fixed in stone** (moving away from single copyrighted measure controlled by publisher).
 - **Jointly owned** (moving away from financial/professional stake in a fixed item/measure).



And, Refinement over Space and Time Means

- Considering questions such as:
 - Does the **conception of quality** vary across contexts?
 - Does the **expression of quality** vary across contexts?
 - If so, are **some conceptions/expressions shared** across contexts (allowing linking)?



Example of an Approach to Iterative Measure Improvement

Many-Facet Rasch Model (MFRM)



Many-Facet Rasch Model (MFRM)

- The **many-facet Rasch model** is one alternative to classical test theory that can contribute to improved measurement.
- Unlike **classical test theory approaches** that tend to focus on **item correlations** and to treat items as **exchangeable...**
- The **MFRM** models the **probability** of a response to an item based on an item's "difficulty" and a classroom's quality "proficiency."
- The locations of the item and classroom on the latent quality continuum are **jointly estimated**.



Many-Facet Rasch Model (MFRM)

- As a Rasch model, the MFRM sits within an **epistemological tradition** that emphasizes iterative measure improvement that can help to address:
 - Rater effects
 - Item variation
 - Standard errors of measurement
 - Among others...
- Other approaches (including the **full suite of IRT models**) have value as well.
 - Each is **complementary**, and other IRT models tend to follow traditions of **expanding the model to fit the data** (including after data collection) rather than **using lack of fit to the model to inform iterative measure improvement**.



Addressing Tensions

- The MFRM has been used in other contexts of **“rater-mediated” measurement** (e.g., college entrance essay exams; Eckes, 2015).
- The MFRM can **support improving reliability and validity** without fully giving up the **humanness of observational ratings**.
- We have been able to use the MFRM in **collaborations**, to **support technical knowledge transfer and open dissemination**.



**Focal Issue #1:
How the MFRM Helps with
Item Variation**



Many-Facet Rasch Model (MFRM)

- **Item variation**
- Think of **array of items** along the latent construct (as in a ruler).
- Encourage **sharp definition of constructs**.
- Encourage writing of **items** to ranging from “**easy**” to “**difficult.**”
- Obtain **empirical item ordering** to test hypothesized item order or inform construct refinement.



Example: SECA

- **Social-Emotional Competency Assessment (SECA)**
 - **Not rater-mediated** (student self-assessment).
 - But good illustration of **basic concepts, as translated for practice.**
- Developed collaboratively with an **IES Researcher-Practitioner Partnership Grant.**
 - **Washoe County School District (Reno, NV)**
 - University of Illinois at Chicago
 - CASEL
- Items are **open source.**
- Technical **knowledge transferred** to district.



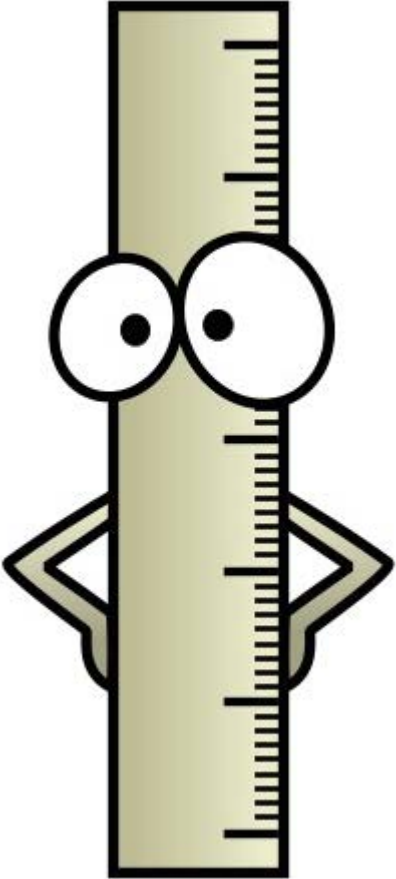
How the District Conveyed the Ruler for Practice

The Rasch Ruler

Measures =
Kids' Levels

Kids who have the MOST
competency

Kids who have the LEAST
competency



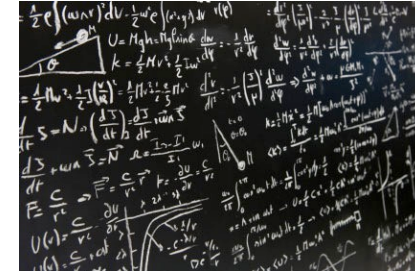
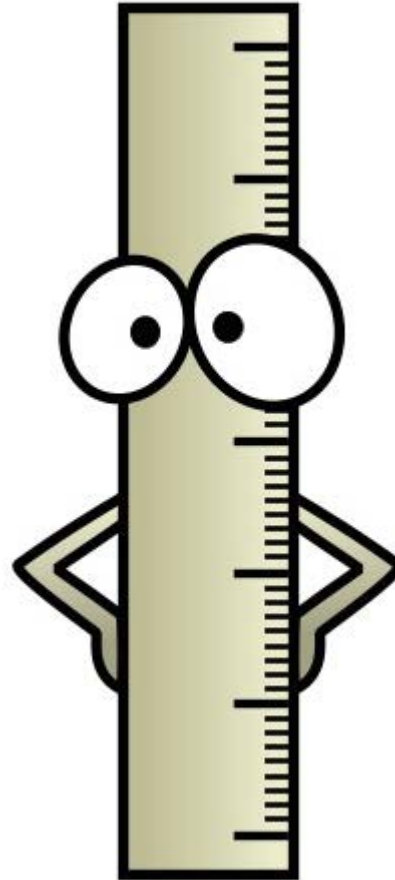
Marks =
Competencies

Competencies that are
really HARD for most
kids

Competencies that are
really EASY for most kids



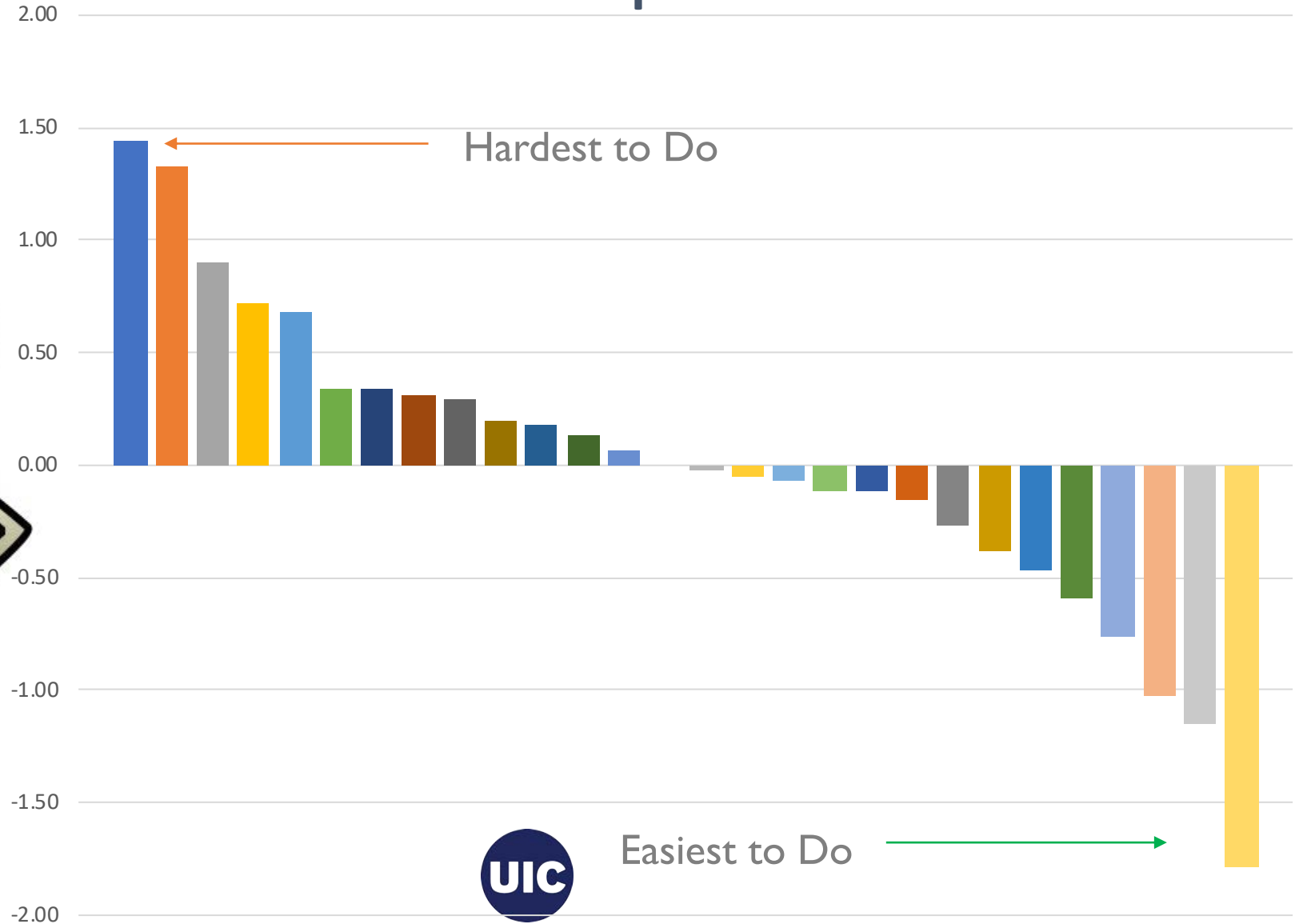
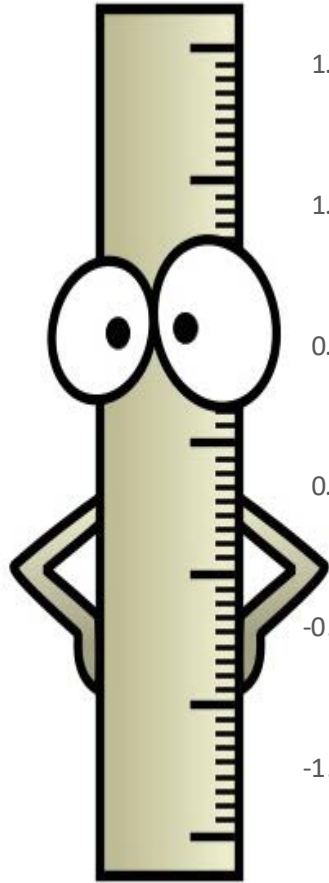
The Rasch Ruler



2+2

If we had marks only at the bottom of the ruler – just the easy math items – we couldn't separate the students with moderately to highly competent math skills.

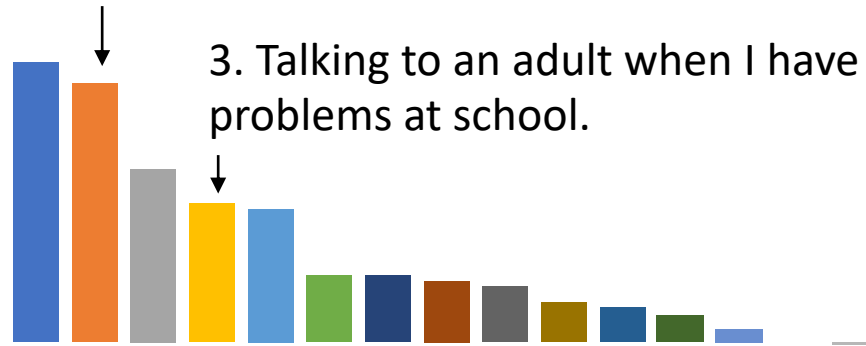
The way WCSD depicted the ruler, showing a well-dispersed set of items: Relationship Skills.



1 = Hardest to Do

6 = Easiest to Do

2. Joining a group I don't usually sit with at lunch.



3. Talking to an adult when I have problems at school.

1. Sharing what I am feeling with others.

5. Getting along with my classmates.

4. Introducing myself to a new student at school.

6. Being polite to adults.



Hypothesizing about Item Order

- We used **state social and emotional learning standards** to develop items and to **hypothesize about expected item order** (grade level).
- We **compared the estimated item locations** to the hypothesized order.

Item	Hypothesized School-Level	Point Estimate	
Social Awareness			
Knowing how to get help when I'm having trouble with a classmate.	Middle	0.29 ^a	Tied for hardest was from a high school standard.
Learning from people with different opinions than me.†	High	0.28 ^a	
Knowing how my actions impact my classmates.	Middle	-0.09 ^b	
Knowing what people may be feeling by the look on their face.†	Elementary	-0.13 ^b	
Knowing when someone needs help.†	Elementary	-0.35 ^c	Easiest was from an elementary standard.



**Focal Issue #2:
How the MFRM Helps with
Precision of Estimation**



Many-Facet Rasch Model (MFRM)

- **Precision of Estimation**
- Estimate location of classrooms and items on the **same scale**.
 - See how **well targeted** the items are at the classrooms.
 - Whether item is easy or difficult is **relative**.
- Items right at classroom's quality level to offer most **information** about that classroom (50:50 chance).
- When items (and their aggregate to the test) have more information, the **standard error of measurement** is lower.
- When the standard error of measurement is lower, the **95% Confidence Intervals** of classroom quality locations are narrower.
- When the 95% CI of classroom quality locations are narrower, classrooms can be **better distinguished** from one another.
- Comparison of such 95% CIs to high stakes cutoffs would be preferable to point estimates to recognize uncertainty of estimation.
- Need numerous items around the cutpoint(s) for narrow 95% CI.



**Focal Issue #3:
How the MFRM Helps with
Rater Effects**



Many-Facet Rasch Model (MFRM)

- **Rater effects**
- Rater locations on **same continuum** as items and classrooms.
- Stable inter-rater (between) effects **can be adjusted**.
 - May reflect pedagogical training, cultural background, etc.
- Additional high resolution insights can **inform manuals and training**.
 - Differential facet functioning.
 - Stable differences *within* raters.
 - e.g., a rater scoring one specific item differently when a teacher is male versus female.
 - Rater fit statistics.
 - Unexpected ratings may reflect idiosyncratic aspects of any rating session.
 - e.g., being tired, feeling hungry, being happy.
- When combined with modern strategies for approximating live **observations with video**, can iteratively rewatch, rescore, improve.



Example: EMOTERS

- **EMOtion TEaching Rating Scale**
 - Content-specific measure.
 - Teaching practices in early childhood classrooms that promote emotion knowledge, expression, and regulation.
- Developed collaboratively with an **IES Measurement Grant**
- Co-PIs: Kate Zinsler (UIC) and Tim Curby (George Mason)
- Co-Is: Rachel Gordon and Cathy Main (UIC)



Aligned to focal child developmental outcomes

Children's Emotion Skills

Domain	Definition	Example Standard	Example Measure
Knowledge	Children's knowledge of cues of their own and others' feelings (facial, physiological, vocal, etc.), their mastery of emotion vocabulary, and their understanding of the causes and consequences of various emotion states.	<ul style="list-style-type: none"> Recognize and label basic emotions (IELDS 20.A.ECa) 	<ul style="list-style-type: none"> Affect Knowledge Test (Denham, 1986)
Expression	Children's abilities to express through face/body/voice how they are feeling and their ability to communicate with others about their emotion needs.	<ul style="list-style-type: none"> Use appropriate communication skills when expressing needs, wants, and feelings (IELDS 20.A.ECb) 	<ul style="list-style-type: none"> Children's Emotion Expression Questionnaire (Halberstadt et al., 1995)
Regulation	Children's abilities to monitor and modify feelings when necessary to meet social expectations and to remain positively engaged.	<ul style="list-style-type: none"> Express feelings that are appropriate to the situation (IELDS 20.A.ECc) 	<ul style="list-style-type: none"> Children's Coping with Negative Emotions Scale (Eisenberg et al., 1993)



Focal teaching practices (based primarily on emotion socialization literature)

Teachers' Emotion Practices			
Domain	Definition	Example Item Content ^a	
		Easier ^b (More frequent)	Harder ^b (Less frequent)
Modeling	Any and all observable adult emotional behaviors and expressions. Such adult emotional displays, whether intentional or not, implicitly teach children which situations are likely to evoke certain emotions, what cues and labels are attached to them, and what reactions and consequences typically follow emotional displays.	<ul style="list-style-type: none"> When T displays negative emotions, fewer versus more children are exposed. T displays positive emotion non-verbally, toward children. T does not display negative emotions non-verbally toward children. 	<ul style="list-style-type: none"> T expresses pretend emotions. T displays positive emotion non-verbally, toward adults. T labels own negative emotion, if expressed.
Responding	Adults' contingent reactions to children's emotion displays, including whether the emotions are validated (labelled, comforted) or invalidated (minimized, punished, ignored).	<ul style="list-style-type: none"> When T has an invalidating reaction, few versus many children are exposed. T reacts positively to a C's negative emotion. T does not react negatively to a C's positive emotion. 	<ul style="list-style-type: none"> T addresses emotion as well as behavior when C exhibits behavior problem. T responds to a C's negative emotion in a validating way. T helps C reduce chance of negative future emotion.
Instructing	Adults' explicit provision of information about emotions, including planned activities, such as following a social-emotional curriculum or choosing an emotion-focused story book to read, or spontaneous moments, such as explaining how someone else is feeling.	<ul style="list-style-type: none"> T asks questions about emotions. T connects emotions to prior events. T labels and demonstrates emotions. 	<ul style="list-style-type: none"> T references her own feelings/emotions. T has C practice new emotion-related skills or knowledge. T provides opportunities for C to share about emotions.
Relating	Closeness, caring, and security of adults' interactions with children, providing the foundation for the emotion climate in a classroom and influencing the extent to which children trust and turn to adults as role models.	<ul style="list-style-type: none"> T joins children in playfulness. T stays physically at C's level during interactions. T seeks opportunities to engage positively with C. 	<ul style="list-style-type: none"> T shares personal information during conversations with C. T shows affection to C during arrivals and departures. C approach T for comfort or affirmation.

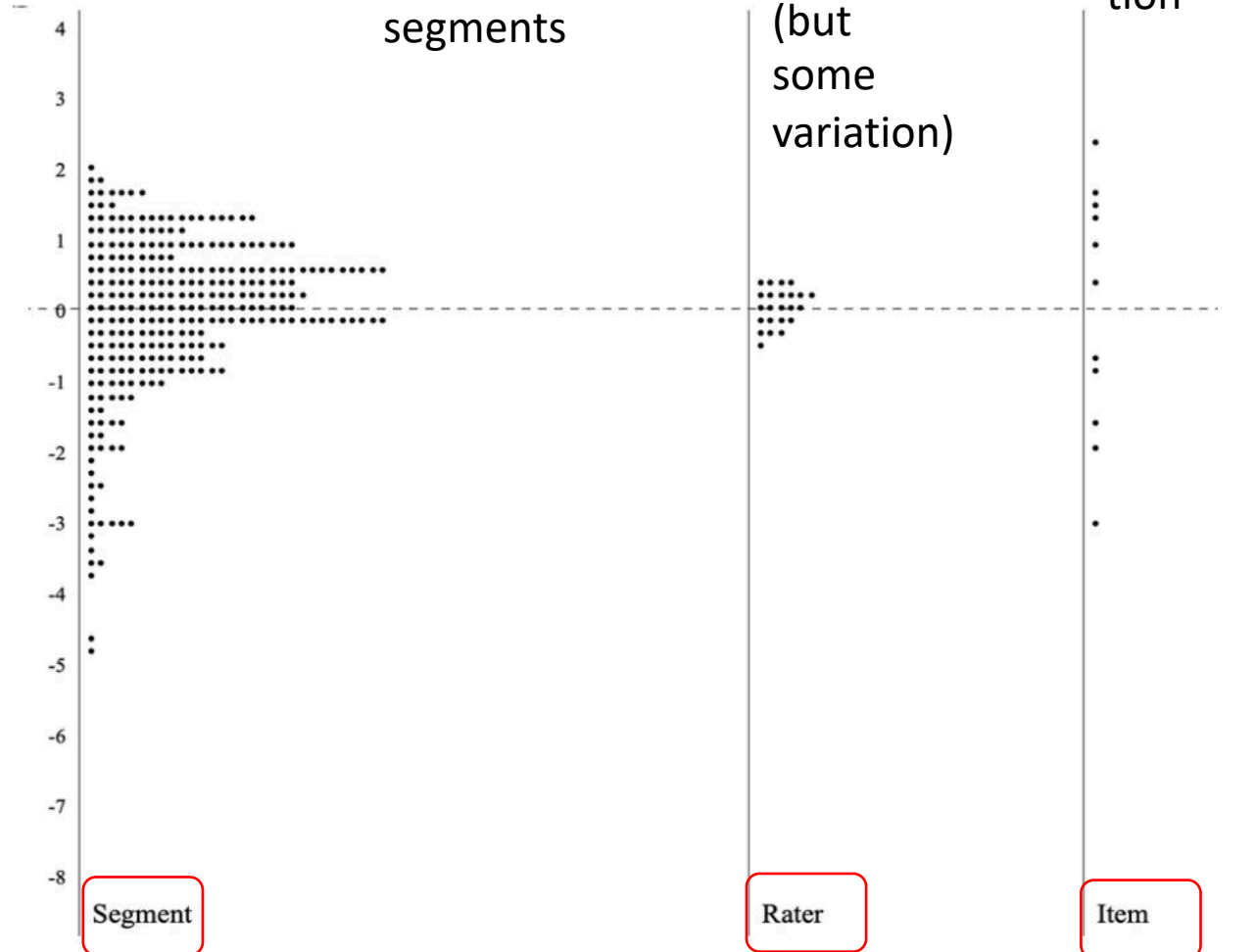
Data for EMOTERS Version 6

- 23 raters
- 18 classrooms
- 1,609 10-minute video segments
- **Multiple cameras** used to approximate what coders would see live in a classroom.
 - panoramic and SWIVL close up
 - SWIVL tracking teacher with 5 cameras around the room.
- **Earlier versions of EMOTERS** used multiple rounds of iterative item development with video from additional classrooms.



Location map for segments, raters, and items

(d) Relating



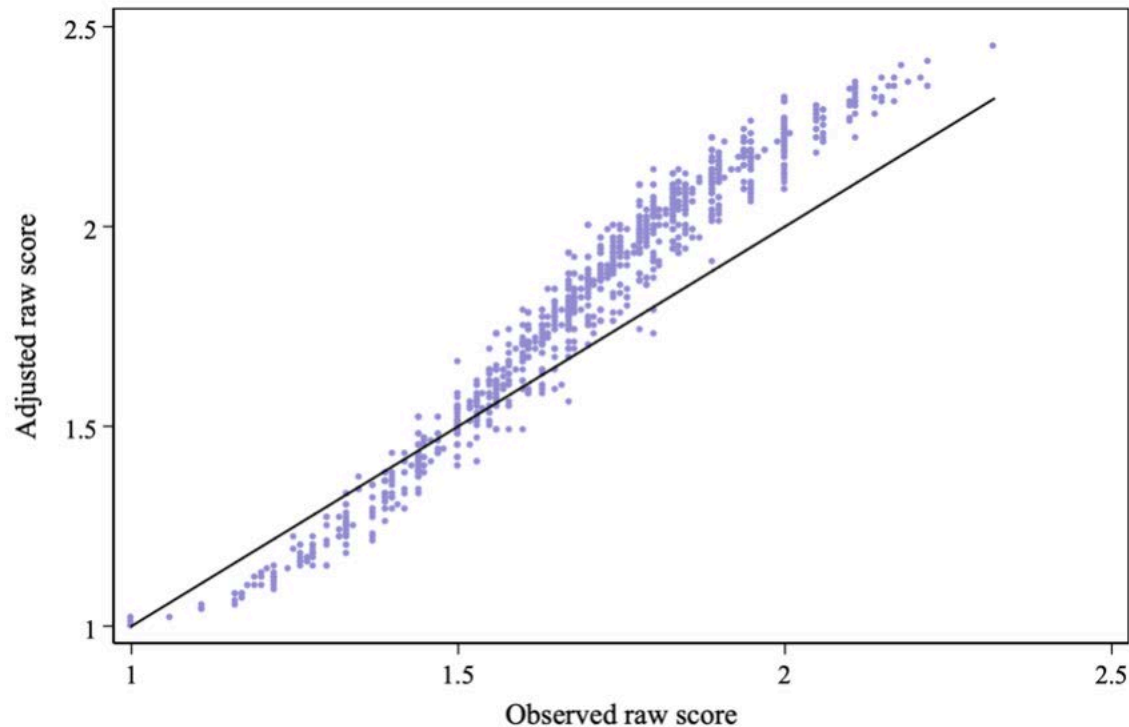
Meaningful substantive item orders

- T shares personal information during conversations with C.
- C approach T for comfort or affirmation.
- T joins C in playfulness.
- T seeks opportunities to engage positively with C.



Segment scores, adjusted for stable inter-rater effects

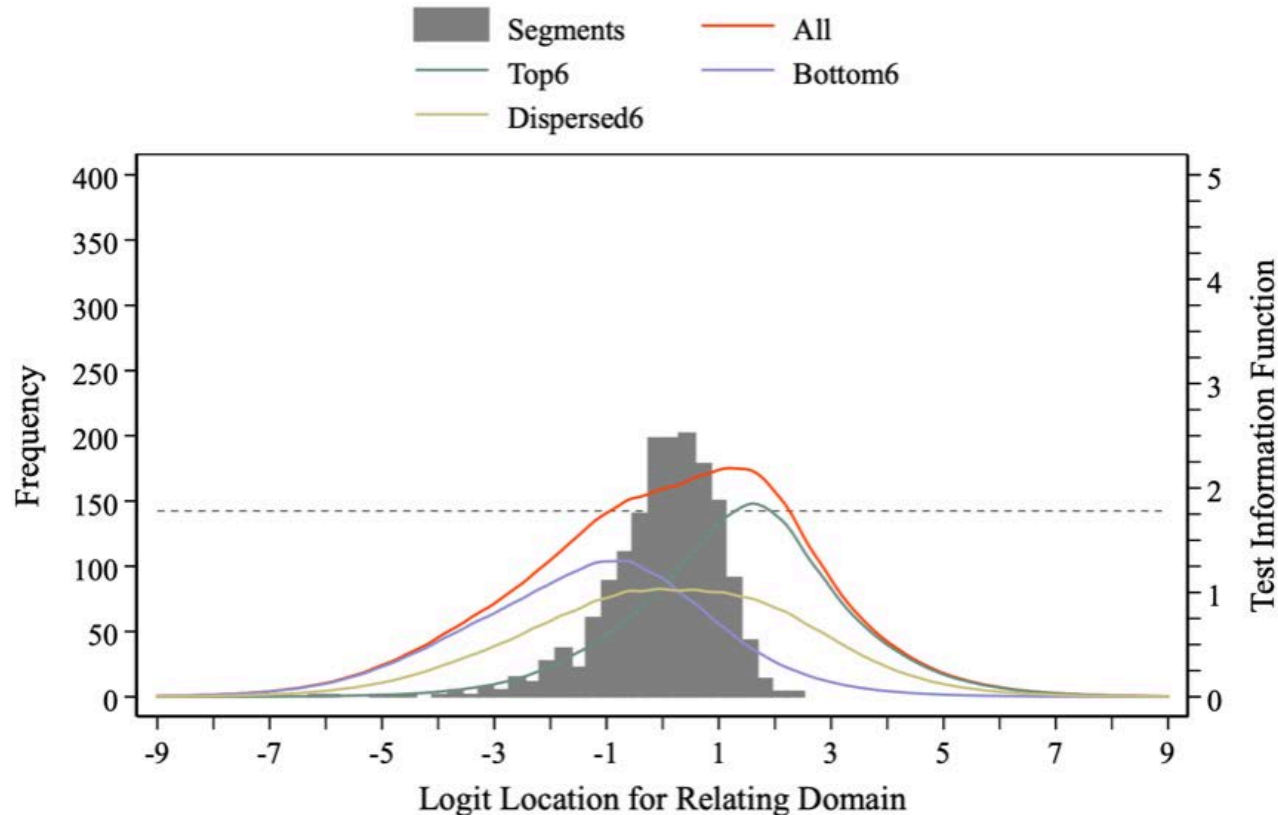
(d) Relating



- Use MFRM to produce adjusted scores (Y axis)
- Plot against raw scores (X axis)
- Line would reflect adjusted scores the same as observed scores.
- Some adjusted upwards, some downwards (reflects raters were located just above and below the middle of segment distribution)
- Adjusted scores are “fairer” since they account for whether a classroom was scored by a rater who tended to be “harsher” or more “lenient.”
- Such adjusted scores would be preferred for research and high stakes uses.

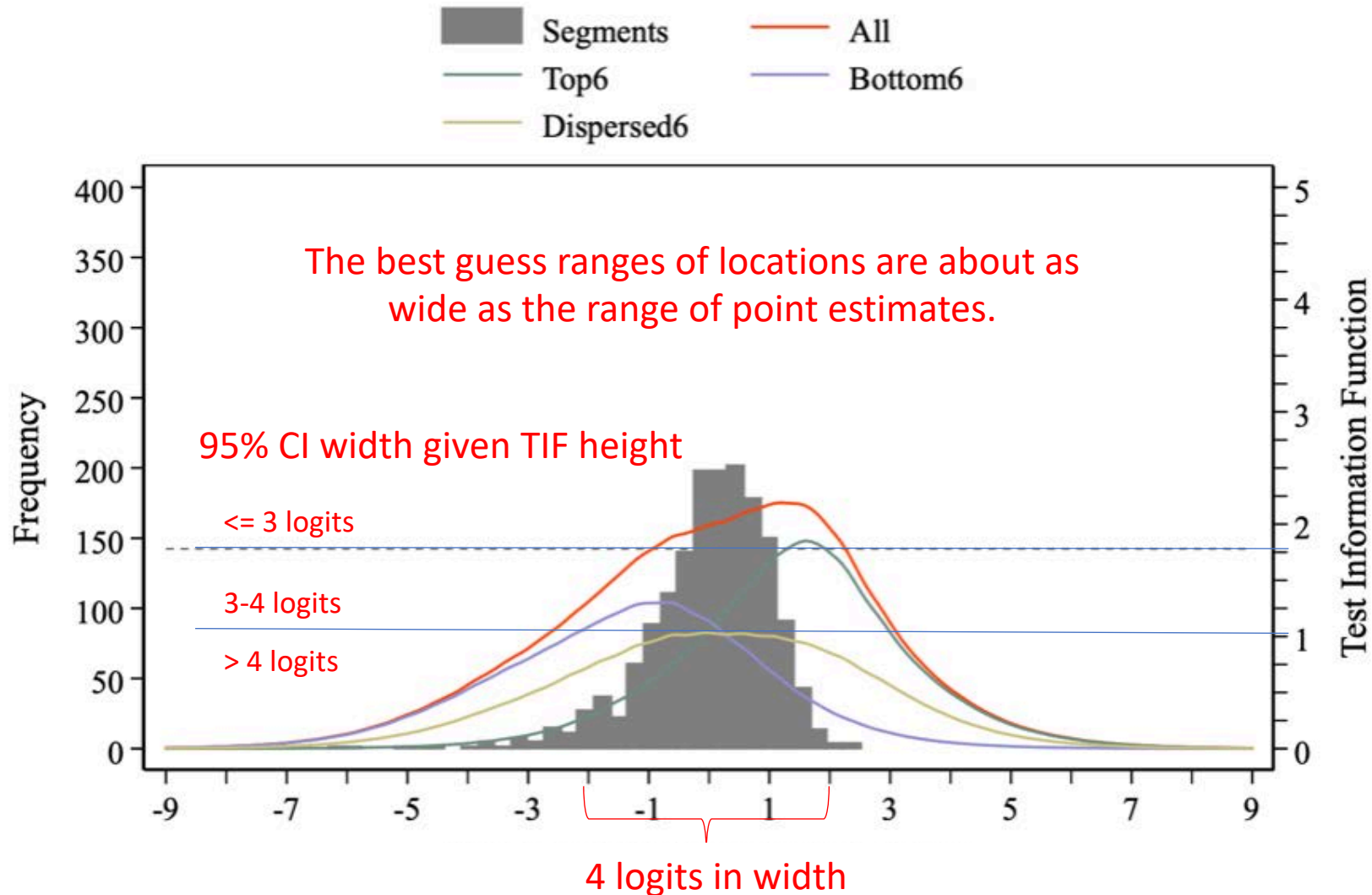
Test information functions (TIFs)

(d) Relating



- Test information functions are shown in lines.
- All-item TIF is red line.
- Sets of fewer (six) items in other colors.
 - Purple items were “easiest” so its TIF peaks in lower region.
 - Green were “harder” so its TIF peaks in higher region.
 - Golden were dispersed across the quality continuum, so similar to but lower than all-items TIF.
- Where they peak, and how high, shows tradeoff of number and locations of items.
- For high stakes cutoffs would want items clustered around the cut score.

(d) Relating Also tell us implications of level of TIF, where segments are located.



Just 30% of pairs of segments differed significantly in locations.

Need more items to increase precision.

Especially need more items in lower range of quality where TIF trails off.

Addressing the Second Tension

- “There is a gap between how psychological science might be optimally conducted and how it is typically conducted, which undermines the credibility of research findings”
- *Kevin King et al. (2019)*
- *Using Implementation Science to Close the Gap Between the Optimal and Typical Practice of Quantitative Methods in Clinical Science.*



Closing the Gap Between Optimal and Typical Practice with Implementation Science

Characteristics of the Method	Individual Characteristics	Inner Setting	Outer Setting
Evidence strength	Peer norms	Implementation climate	External policies
Relative advantage	Knowledge	Culture	Rewards
Adaptability	Beliefs	Available resources	Incentives
User-friendly design	Self-efficacy	Rewards	Culture
Complexity	Behavioral control	Incentives	
Cost	Needs		
	Resources		

Need to intervene at each level:

- Making method easier to adopt.
- Promoting user beliefs and capacity.
- Supporting local culture.
- Supporting professional culture.

And monitor what works (and what doesn't work).

Source: King et al. 2019. Journal of Abnormal Psychology.



Addressing the Third Tension



FEDERAL REGISTER

The Daily Journal of the United States Government



Ⓜ Rule

Open Licensing Requirement for Competitive Grant Programs

A Rule by the [Education Department](#) on 01/19/2017



SUMMARY:

The Secretary amends the regulations of the Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards in order to require, subject to certain categorical exceptions and case-by-case exceptions, that Department grantees awarded competitive grant funds openly license to the public copyrightable grant deliverables created with Department grant funds.

Addressing the Third Tension (cont.)

- **Arguments for closed source**
 - **Control fidelity** of implementation
 - **Incentivize innovation** (through monetary return)
 - **Fund dissemination**
- **Experience** with current high stakes use suggests
 - **Certification standards** did not ensure adequate reliability
 - Measures **did not quickly improve** as evidence accumulated
 - Policy on **“reliability and validity”** may promote static thinking
 - **Copyright** reduces adaptability
 - Choice of single/few measures **ensured monopoly**



Addressing the Third Tension (cont.)

- **Arguments for open source**
 - Promotes adaptability
 - Avoids “recreating wheel”
 - Allows public return for taxpayer investment
 - Promotes public peer review
 - Reduces cost and increases access
- **Potential models to consider include**
 - **Completely open** dissemination
 - Dissemination with **CC-BYNCSA**
 - Attribute authorship
 - Prevent commercialization
 - Share alike (if remix)
 - **Transparency of cost structures**
 - Are fees covering costs, reinvested in development, or making profit?
 - Are visible companies nonprofit, but subsidiaries for profit?
 - **Transparency of evidence**
 - Are data gathered through publicly funded use available for re-analysis?
 - Are those tested viewed as co-owning their responses?
 - Are those observed for measure development viewed as co-owning their practices?
 - Are stakeholders engaged in study design and interpretation?



Acknowledgments

- This work draws primarily from collaborative examinations of the psychometric properties of measures of classroom quality and children's socio-emotional development in early childhood funded by IES and NIH:
 - IES R305A090065
 - IES R305A130118
 - IES R305A160010
 - IES R305H130012
 - NIH R01HD060711
 - NIH R03HD098310
- Results reflect our teams' interpretation (not necessarily those of our funders).
- Presentation reflects my synthesis (not necessarily individual team members).



QUESTIONS

Rachel A. Gordon
ragordon@uic.edu

College of Liberal Arts and Sciences | Department of Sociology

