

# High Stakes Quality Measures in Early Care and Education: Reconsidering the Evidence

Rachel A. Gordon, Professor  
Sociology and Institute of  
Government and Public Affairs  
University of Illinois at Chicago

***The Challenge of Using Observational Quality  
Rating Tools in Accountability Systems and  
Strategies to Address Them***

*National Research Conference on Early Childhood  
July 13, 2016 (v3)*

# 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**
  - **NIH R01HDo60711**
  - **IES R305A130118**
  - **IES R305A160010**
  - **IES R305H130012**
- Results reflect our teams' interpretation (not necessarily those of our funders).
- Presentation reflects my synthesis (not necessarily individual team members).

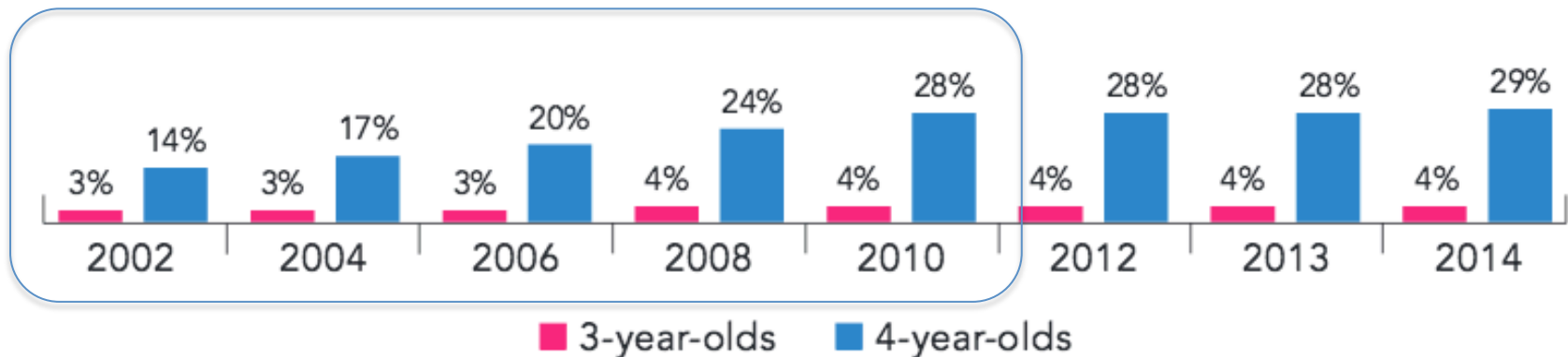
# Preview of Talk

- Brief reminder of policy context, especially high stakes use of measures.
- Highlights from research findings.
  - General importance of freshly considering the evidence base specifically for each use of a measure.
    - High stakes, professional development, research, self assessment.
  - Current measures in high stakes use.
    - Do they predict large school readiness gains?
    - Do they sharply measure constructs aligned with readiness gains?
    - Are they constructed for maximal precision (high signal vs. noise)?

**Brief Reminder:**  
Public Investments and  
High Stakes Use of Measures

# Expanding Public Investments in Early Care and Education

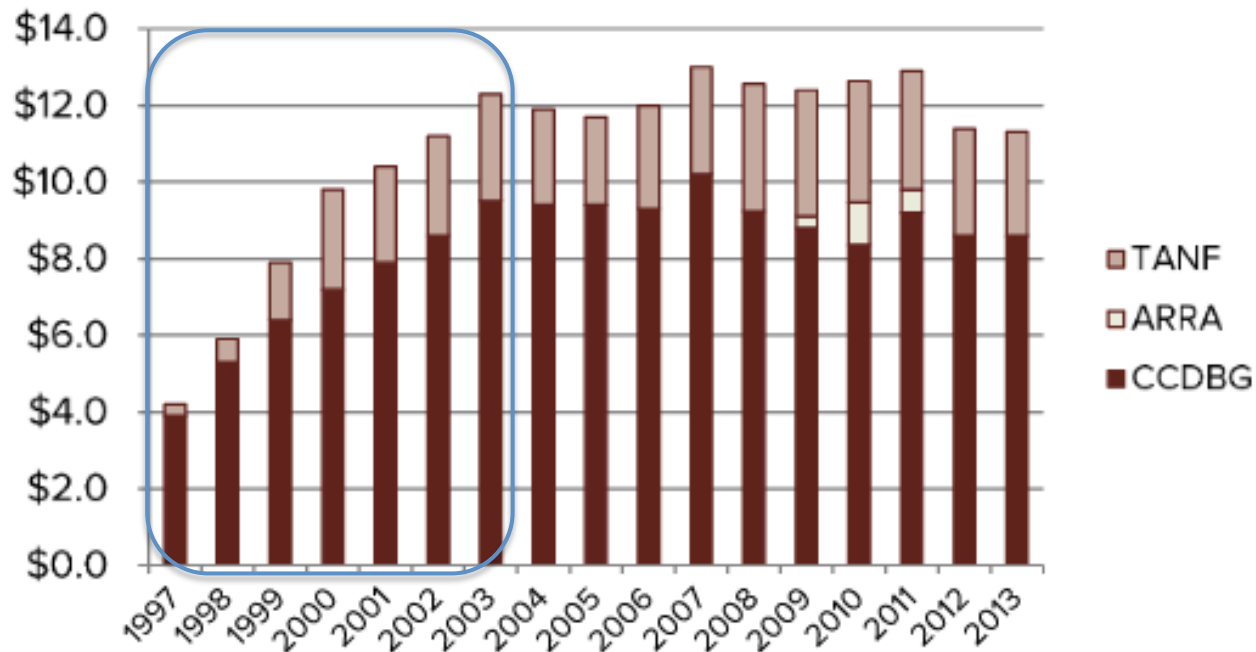
NIEER: State-Funded Prek Enrollment  
PERCENT OF NATIONAL POPULATION ENROLLED



Percentage of 4 year olds enrolled in state pre-k doubled, 2002 to 2010

# 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 tripled, 1997 to 2003

# Policy Focus on *Quality* Early Care and Education

- 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**.

# 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.

# 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](mailto: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.

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. Ideally, consider:
  - What are the policy *goals*?
  - What aspects of quality *align* with these goals?
  - What are the *best strategies* for assessing these aspects of quality for this *particular use*?
- As the desire to assure quality in publicly funded programs grew rapidly, decision makers turned to existing measures.

# ECERS-R and CLASS

- Two observational measures most widely used: ECERS-R and CLASS.
- 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.

# 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 scores assigned based on observed 400+ indicators.
- New version: ECERS-3.

# ECERS-R Item 10: Meals/Snacks

Inadequate 1	2	Minimal 3	4	Good 5	6	Excellent 7
<b>10. Meals/snacks</b>						
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>serving</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/disposing 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).		
		3.4 Nonpunitive atmosphere during meals/snacks.		5.4 Dietary restrictions of families followed. <i>NA permitted.</i>	→	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).		3.5 Allergies posted and food/beverage substitutions made. <i>NA permitted.</i>				
1.5 No accommodations made for children's food allergies. <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.

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

# CLASS

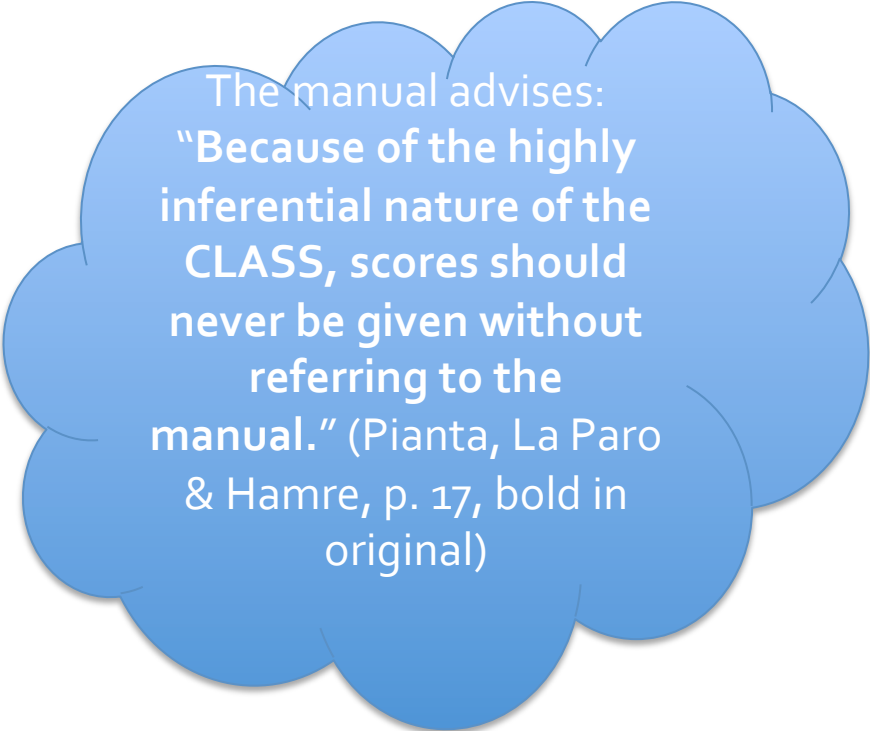
- Developed in *1990s/2000s* beginning in a *research* study and later aimed at *professional development and coaching*.
- Reflects *developmental theory and research* and emphasizes teacher-student (adult-child) *interactions* as the primary mechanism of development and learning.
- Observers assimilate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

### **Low Quality of Feedback (1, 2)**

---

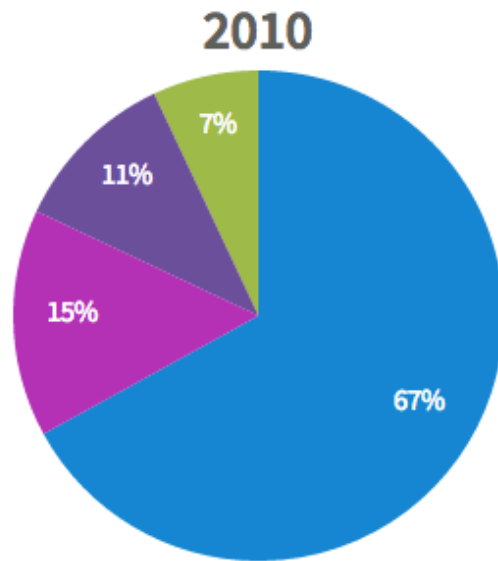
***The teacher rarely provides scaffolding to students but rather dismisses responses or actions as incorrect or ignores problems in understanding.*** In scaffolding, a teacher acknowledges where a student is starting and provides the necessary level of help to allow the student to succeed or complete 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 and a big 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.

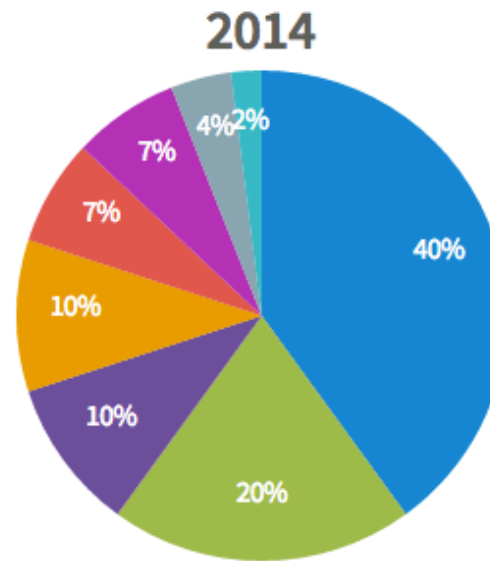
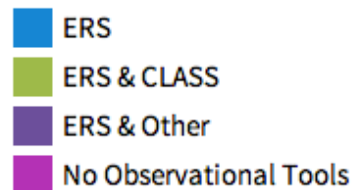


The manual advises:  
"Because of the highly  
inferential nature of the  
CLASS, scores should  
never be given without  
referring to the  
manual." (Pianta, La Paro  
& Hamre, p. 17, bold in  
original)

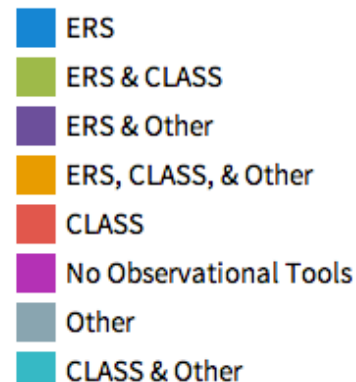
# Use in State Quality Rating and Improvement Systems



Most commonly used observational tools in 2010.



Most commonly used observational tools in 2014.



## 2010

~ 85% ECERS-R  
~ 7% CLASS

## 2014:

~ 87% ECERS-R  
~ 37% CLASS

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



# Example: Illinois' QRIS Learning Environment

Program demonstrates high quality of classroom environment

## LICENSED CHILD CARE CENTER

ERS<sup>1</sup> average overall score:  
At least 4.5 with no classroom  
below 4.0, verified by on-site  
assessment by state-approved  
assessor

*OR*

CLASS<sup>2</sup> Emotional Support and  
Classroom Organization:  
average scores above 5.0 with  
no classroom below 4.0 as  
verified by on-site assessment by  
state-approved assessor<sup>3</sup>

*OR*

Accredited sites: Evidence from  
state-approved accrediting body

## PRESCHOOL FOR ALL

ERS<sup>1</sup> average overall score:  
At least 4.5 with no classroom  
below a 4.0, verified by on-site  
assessment by state-approved  
assessor

*OR*

CLASS<sup>2</sup> Emotional Support and  
Classroom Organization:  
average scores above 5.0  
with no classroom below 4.0  
verified by on-site assessment  
by state-approved assessor

## HEAD START/ EARLY HEAD START

ERS<sup>1</sup> average overall score:  
At least 4.5 with no classroom  
below a 4.0, verified by on-site  
assessment by Head Start  
approved assessor

*OR*

CLASS<sup>2</sup> Emotional Support and  
Classroom Organization: average  
scores above 5.0 with no  
classroom below 4.0 verified  
by on-site assessment by Head  
Start approved assessor

*AND*

Head Start Program Performance  
Standards are in compliance:  
1304.21(a)(1), 1304.21(a)(3).

# Use in Head Start

## What do the Head Start CLASS® review scores mean?

### What CLASS scores cause a grantee to be required to compete?

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.

# Highlights:

## Evidence for High Stakes Use

*I am briefly highlighting findings in the interest of time.  
I have listed citations, and would be happy to share full  
publications or additional details.*

# Evidence for High Stakes Use

- Alternative options for high stakes use.
  - 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?
- Potential for limits of ECERS-R and CLASS based on absolute level of evidence.
  - Both were designed for purposes different from their current high stakes use.
  - Both were embedded in practice/ professional and conceptual/empirical knowledge.
  - But, neither used a fully modern measurement development and psychometric approach (e.g., IRT) during development, especially one aligning with this particular high stakes use.

Do ECERS-R and CLASS predict large  
school readiness gains?

# Accumulating Evidence: Small Associations (Often Not Sig.)

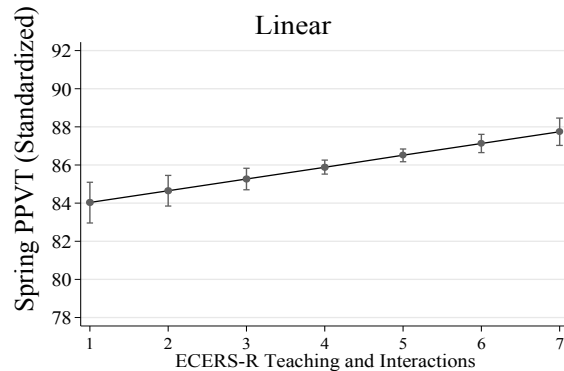
- Burchinal, Kainz and Cai (2011)
  - Effect sizes (adjusted correlations) of *.14 and below* across published studies with a range of child outcomes.
  - *Even* when focusing on *low-income* children and *aligning* subscales with language, math, social, and behavioral outcomes in new analyses, 32 of 36 adjusted correlations at or below .10.
- Keys and colleagues (2013)
  - Average effect sizes between *.01 to .05* for language, math, social, and behavioral outcomes.
- Some evidence of thresholds (stronger effects in higher quality region; Burchinal, Zaslow, & Tarullo, 2016).
  - But still *small* for ECERS-R and CLASS.
  - And demonstrated *small sample sizes* across *regions of quality* and need for data collected specifically to test this question (oversample higher quality).

# New Data Syntheses (IES R305A130118)

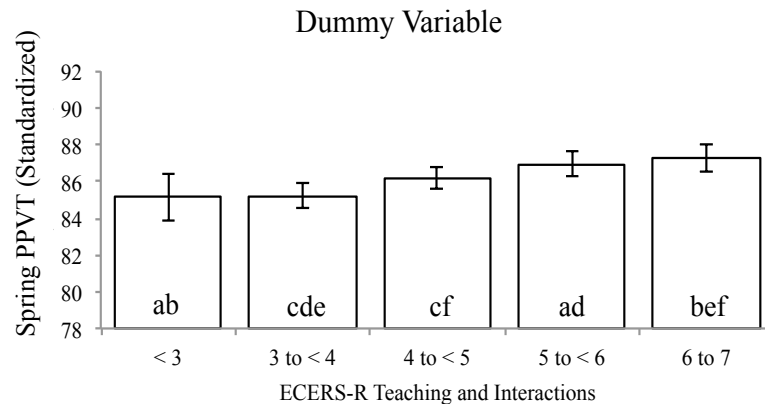
- Meta-analyses
  - 13 datasets with multiple waves and subgroups
- Integrative data analysis (stacked datasets)
  - FACES 2000, 2003, 2006 and 2009
  - Greater sample sizes across the quality continuum
- Used numerous strategies to examine non-linearity (dummy variable, quadratic, piecewise, non-parametric).
- Used numerous strategies to examine sensitivity (e.g., multiple approaches to missing data, complex samples, quality and outcome scoring).

# Example: Predicting Growth in Children's Vocabulary (PPVT)

## ECERS-R Teaching and Interactions

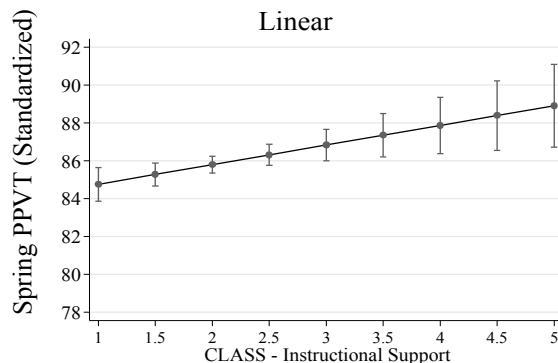


Effect  
size:  
.05

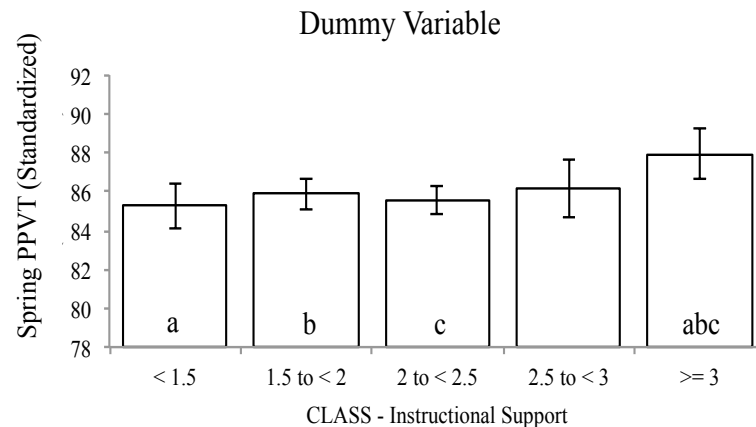


Max  
Std.  
Diff:  
.13

## CLASS Instructional Support

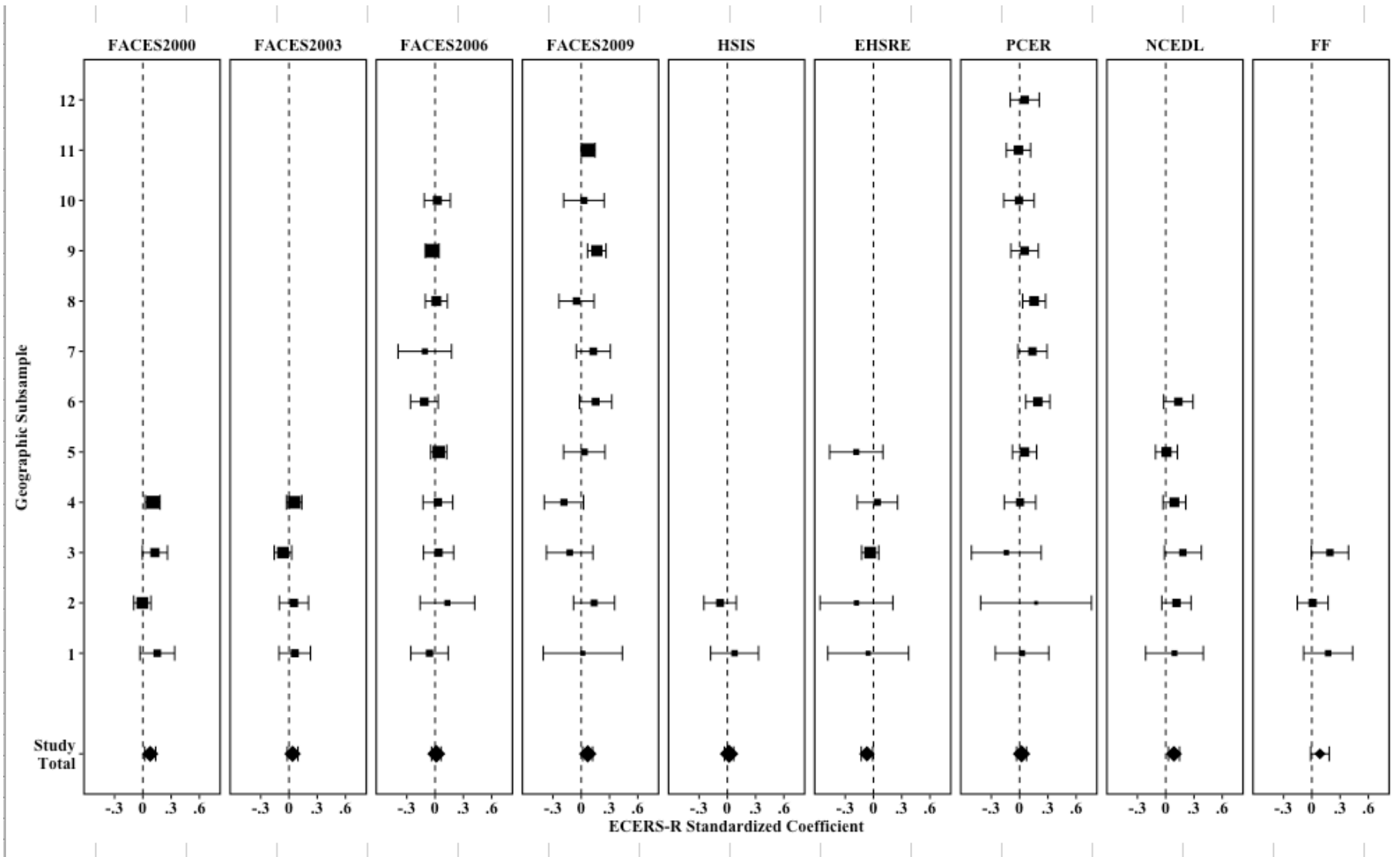


Effect  
size:  
.04



Max  
Std.  
Diff:  
.17

# Example: Predicting Growth in Children's Vocabulary (PPVT)



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

# Importance of Dimensions of Quality

- Ideally, high stakes measures would be created specifically to measure aspects of quality aligned with policy goals.
  - Content-focused aspects of quality aligned with particular readiness domains may show stronger relationships.
- Alternatively, if measures designed for other purposes are used, they should have clear definitions of the aspects of quality measured and empirical evidence that items measure them.
  - Yet, evidence for the subscales defined in ECERS-R and CLASS manuals (and often used in policy) is limited.

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

- The ERS presumes 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” <http://ers.fpg.unc.edu/>
  - 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, and the most aligned of these are often somewhat more highly correlated with outcomes; these dimensions are *sometimes used in QRIS*.

# Two Dimensions Replicate (IES R305A130118)

- To solidify evidence related to dimensionality, and encourage consistent practice/policy use, we factor-analyzed data from eight surveys (with 14 waves) and synthesized the results.
- Two broad dimensions replicated across the datasets:
  - Language-Reasoning/Interaction  
(LR: Items 16-18; Int: Items 29-33).
  - Space Furnishings/Activities  
(SF: Items 2-6; Act: Items 19-26 & Item 28).
- But associations with outcomes still small.

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

- CLASS PreK manual produces scores in *three broad domains*: Emotional Support, Classroom Organization, Instructional Support. (<http://teachstone.com/>)
- Instructional Support captures dimensions of teacher practice meant to *cut across content* (Concept Development, Quality of Feedback, Language Modeling).
- Dimensions sometimes more strongly related to outcomes when aligned by domain (but still small).

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

- CLASS developers recently published a different “*bi-factor*” structure for the CLASS PreK (Hamre, Hatfield, Pianta & Jamil, 2014) that *differs* from the subscales written into policy.
  - *One general* dimension (responsive teaching).
  - *Two specific* dimensions (proactive management and routines; cognitive facilitation).
  - Somewhat more conceptually consistent pattern with outcomes, although effect sizes *still small* (<.10).
- We replicated this bi-factor structure, although like the CLASS team had problems with *convergence* (perhaps due to item skewness and correlation).

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” so not all were rated.
  - CLASS a highly inferential approach, where coders assimilated all they’ve seen in their heads (rather than in checklists).

# ECERS-R Standard

## “Stop Scoring”

- Conditions in the indicators of lower scores must be met before indicators of higher scores are evaluated.
- Higher score may not always reflect higher quality, especially for aspects of quality relevant for school readiness.

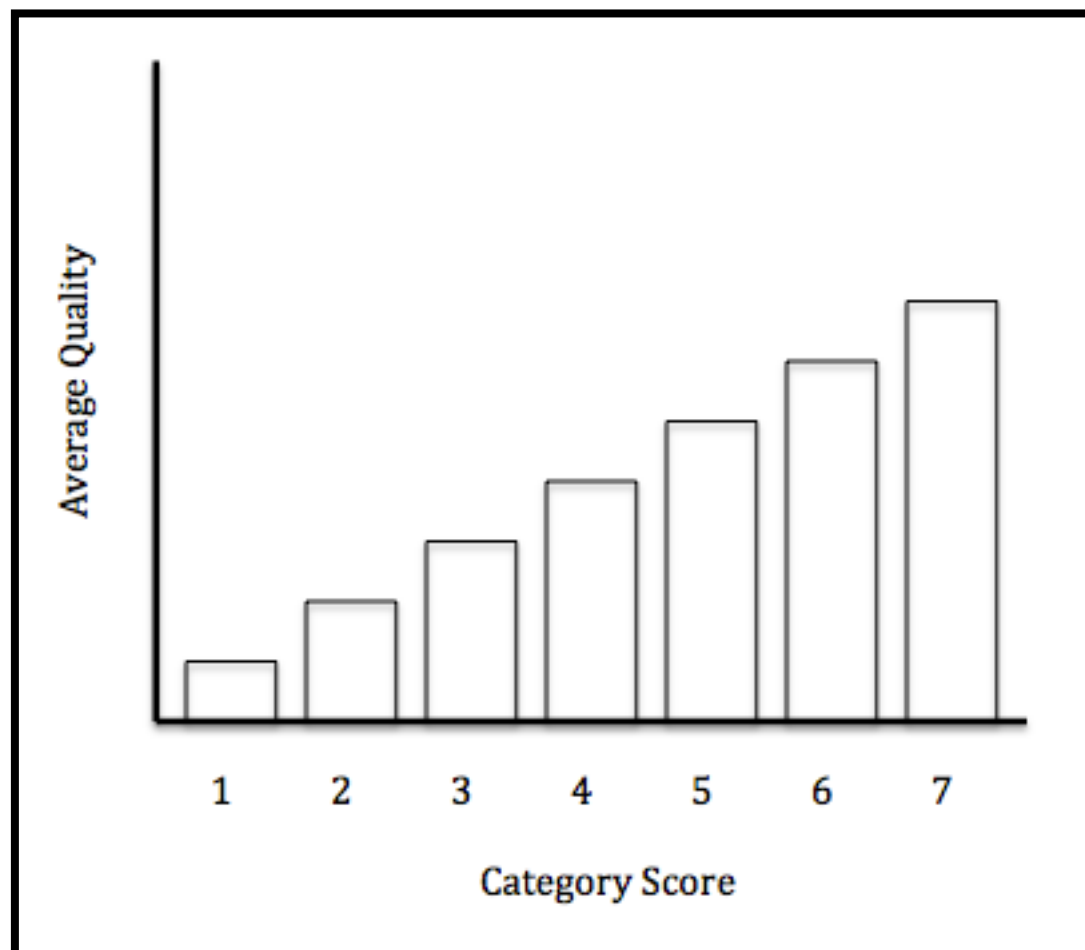
# ECERS-R Item 10: Meals/Snacks

Inadequate 1	2	Minimal 3	4	Good 5	6	Excellent 7
<b>10. Meals/snacks</b>						
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>serving</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 Would Order Look Like?

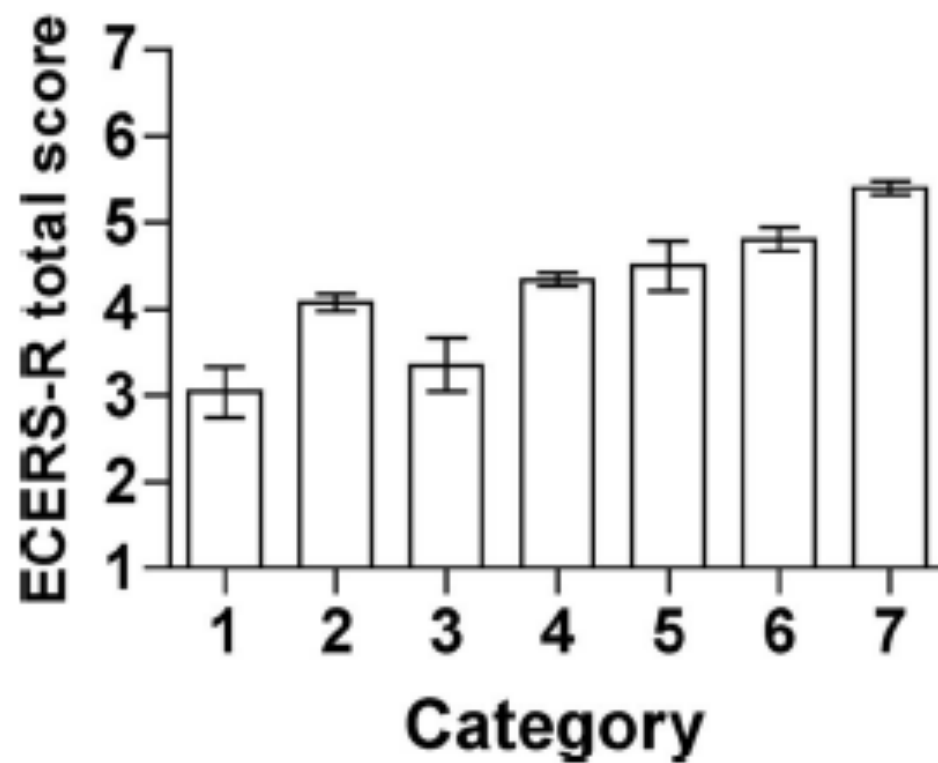
- If higher scores reflect higher quality, then average quality scores should be higher for centers rated in higher categories versus lower categories.



# What Would Order Look Like?

- If higher scores reflect higher quality, then average quality scores should be higher for centers rated in higher categories versus lower categories.
- Alternatively, may see unexpected flat regions or dips in average quality at some higher category scores.

(c) ECERS-R Total Score



# Non-Order in Category Averages (IES R305A130118)

- Category averages out of order for some items in all 8 datasets.
- In analysis of stacked data files (with greatest precision) over two-thirds of items had at least one pair of adjacent category means that did not progress in a stair step fashion.
- Most common location of non-order was categories 2-to-3 followed by categories 4-to-5.
- The problem was most evident in the Personal Care Routines items and least evident in the Language-Reasoning items.
- For nearly  $\frac{3}{4}$  (26 of 36) items, at least one category-total point biserial correlations was negative.

# IRT Models Also Identify Disorder (IES R305A130118)

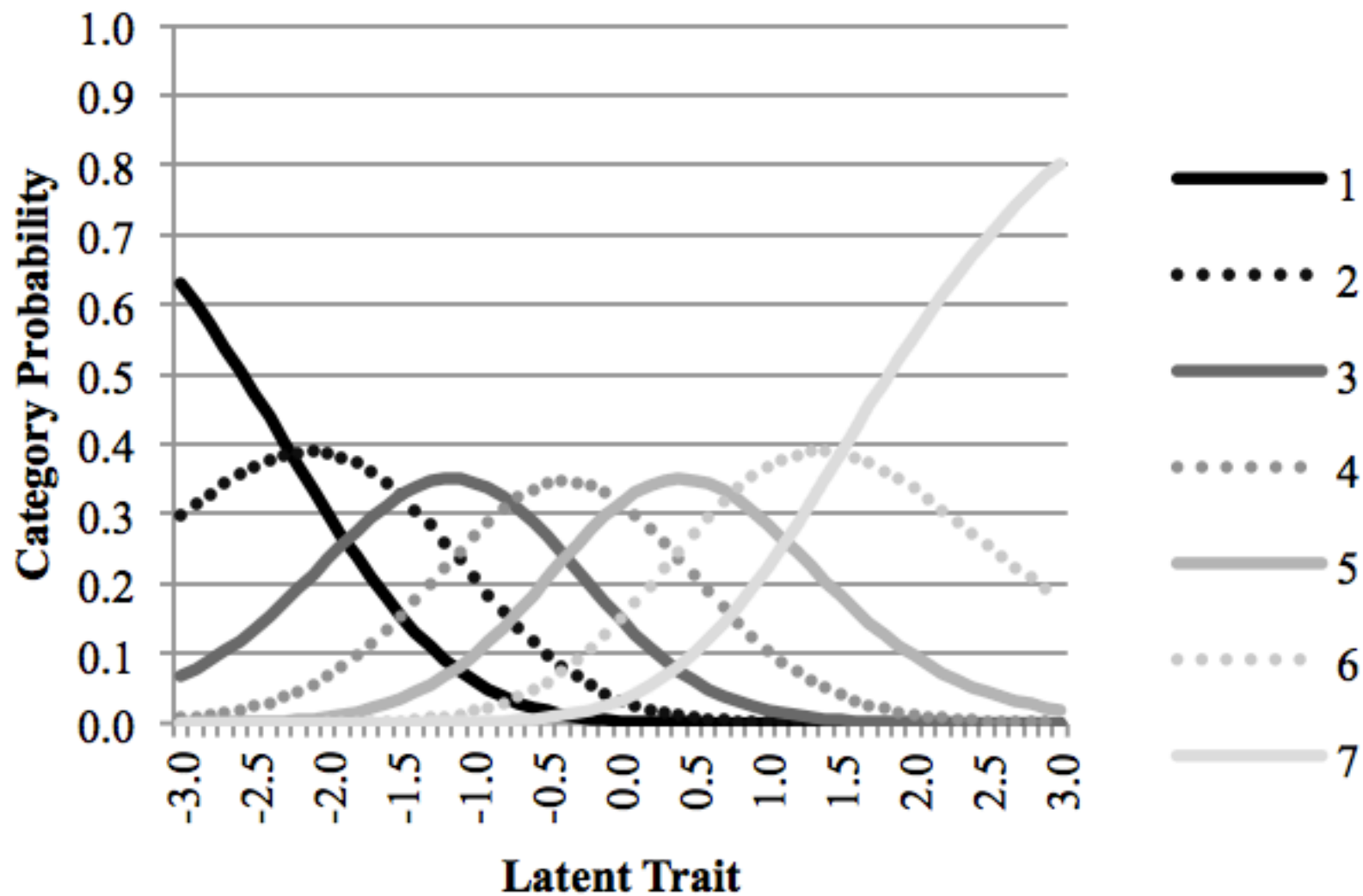
- Partial Credit Model

- For every item, at least one pair of adjacent thresholds - latent level of quality where a rater is equally likely to choose between adjacent categories – was out of order.
- This disorder generally involved the 3<sup>rd</sup> and 5<sup>th</sup> categories.

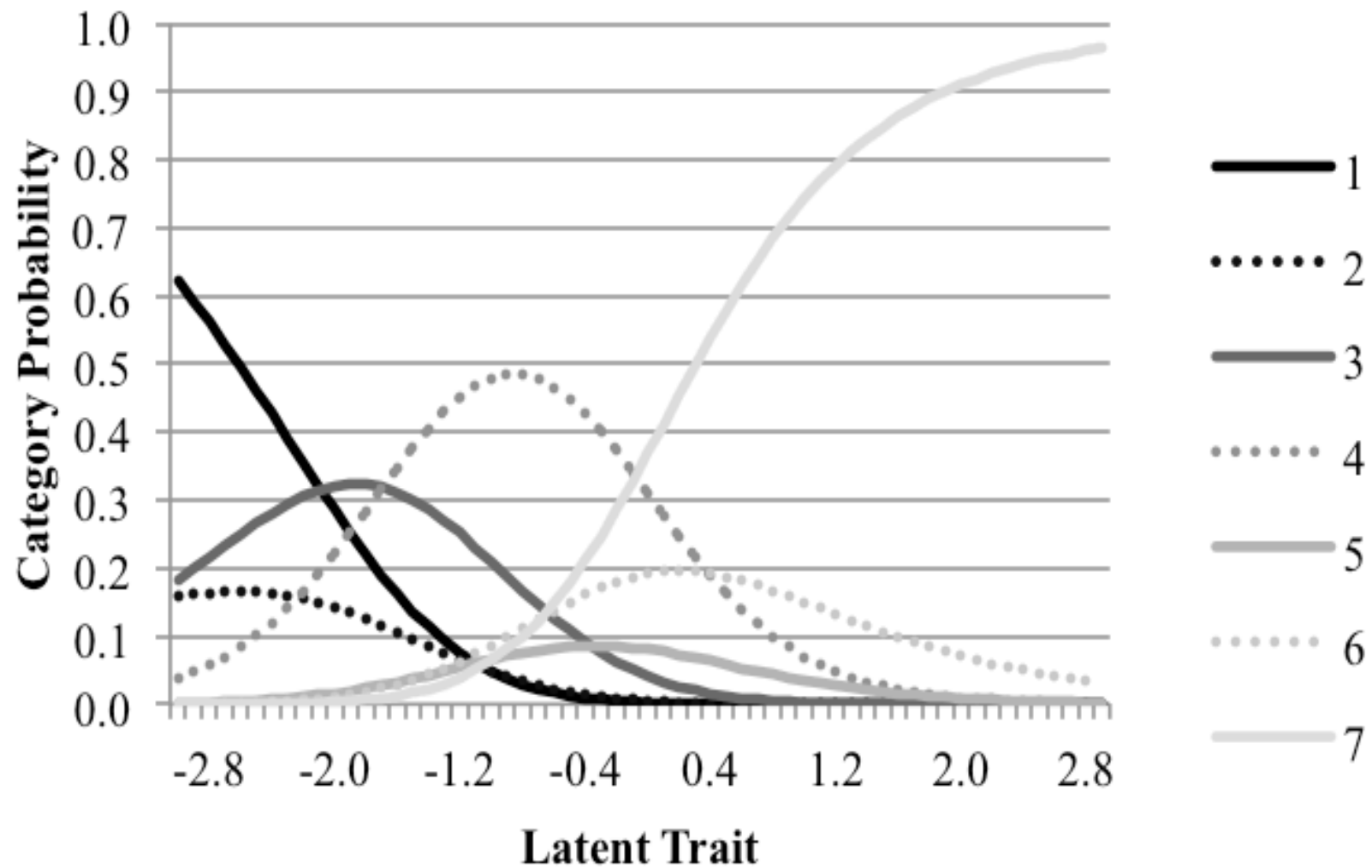
- Nominal Response Model

- Category boundary discriminations negative (and scoring function values out of order) in at least one place for 15 of 36 items (42%).
- Category boundary discriminations small (0 to 0.5) (and scoring function values progressed minimally) for 35 of 36 items.
- This nonorder typically occurred between scores 2 and 3 and scores 4 and 5.

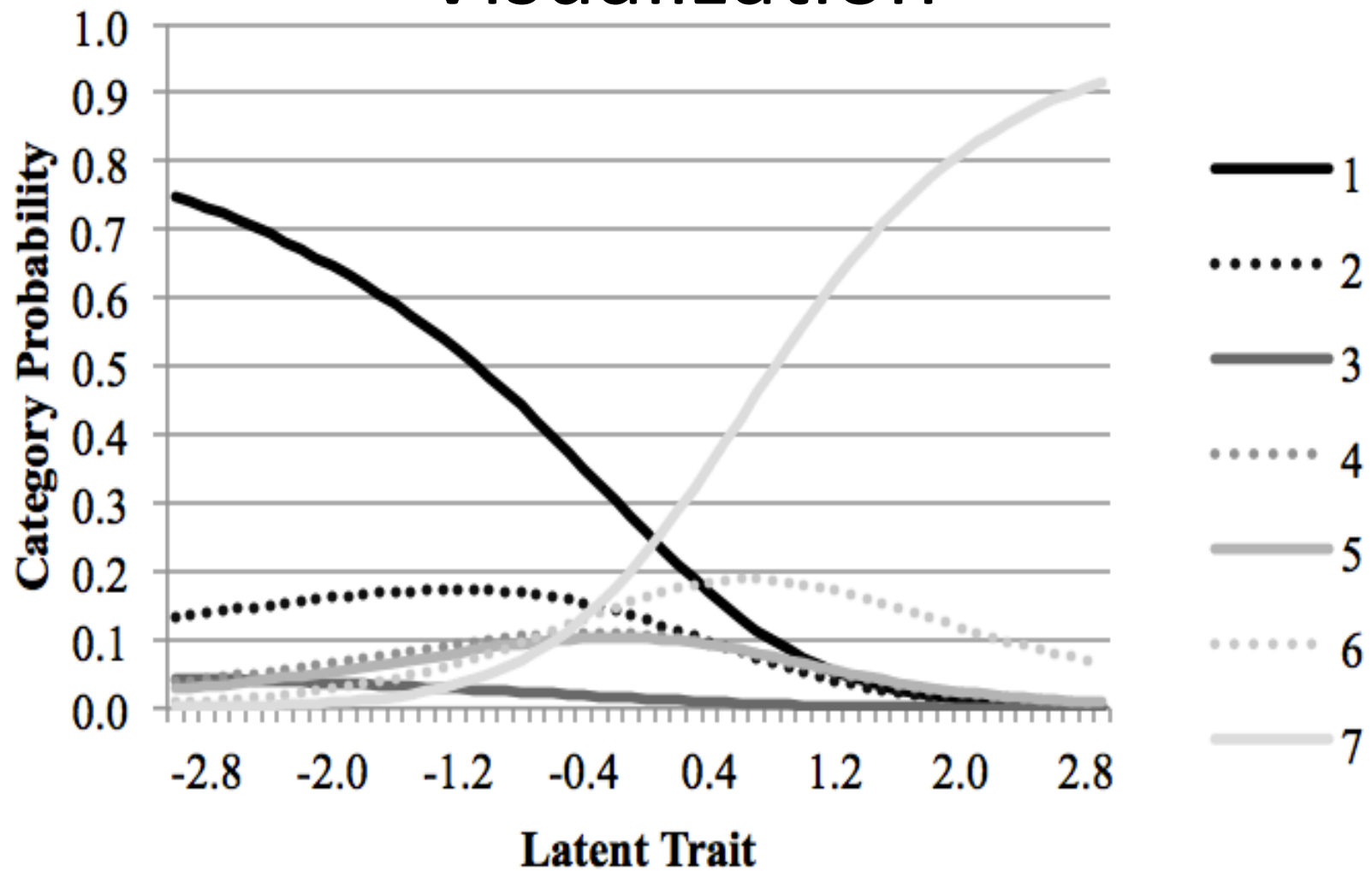
# Visualization



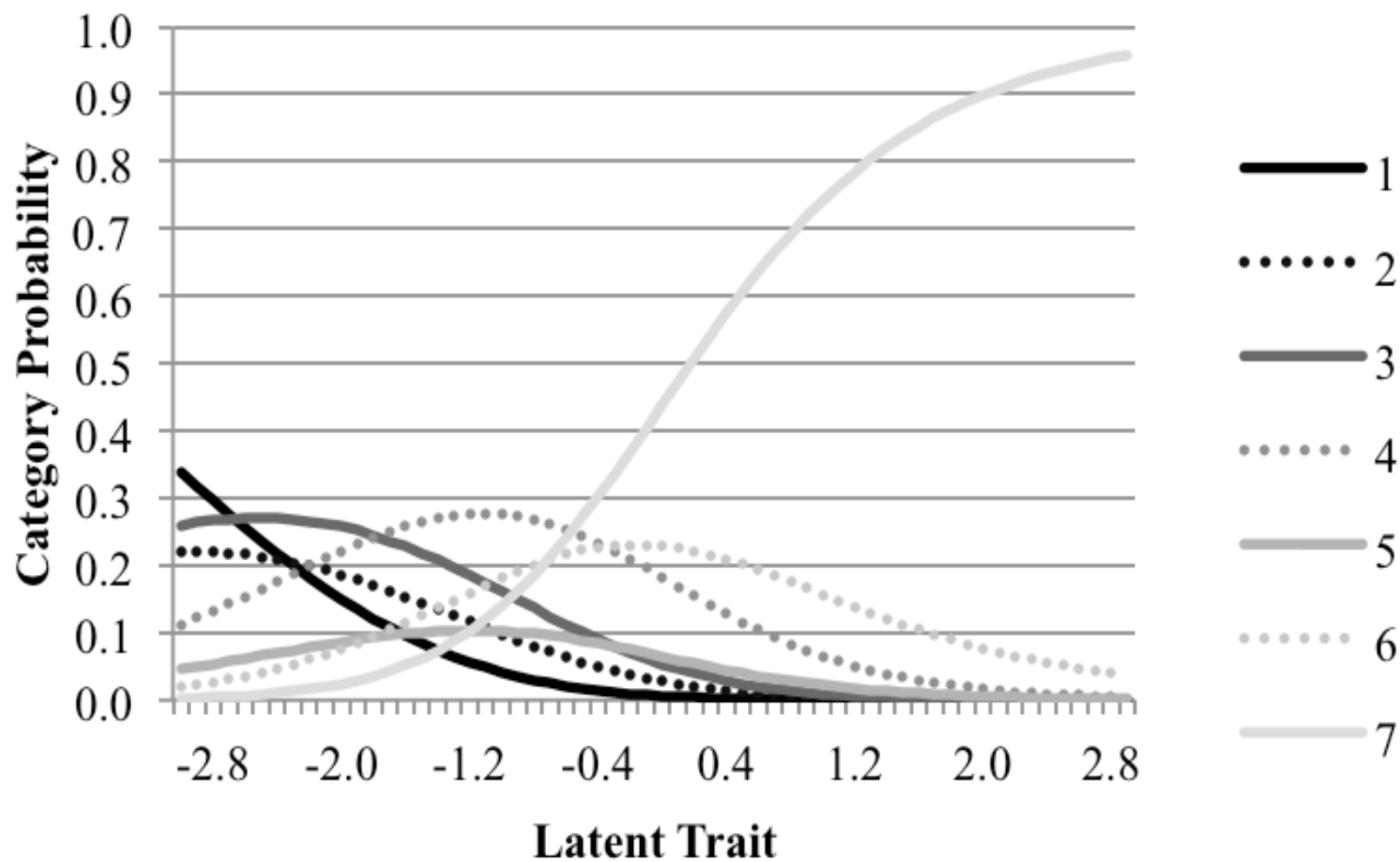
# Visualization



# Visualization



# Visualization



# Interpretations

- Consistent location may reflect different scoring rules for even and odd categories (requiring half vs all indicators to be met).
- Greatest problems evident for personal care routines may reflect mixing of different aspects of quality (also indicator analysis in Gordon et al., EED, 2015).

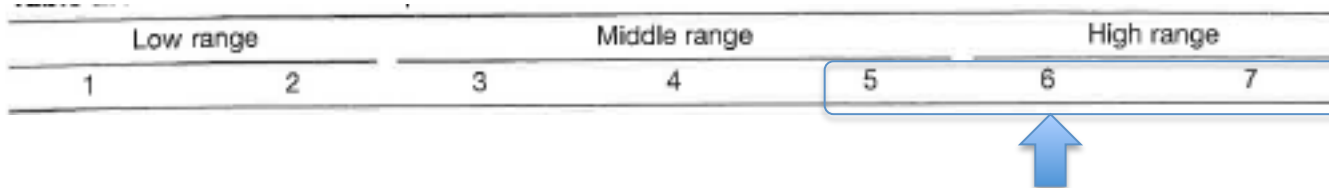
# Implications

- *These results caution use of the simple sum (average) of the raw scores, including in relation to high stakes cutoffs.*
- Preston and Reise (2015, p. 392)
  - When CBD values are not positive, a response in a higher category “does not indicate more of the trait” than a response in a lower category.
  - “when category distinctions fail to discriminate, a researcher would not want to use a scoring strategy that aggregates raw integer item scores.”

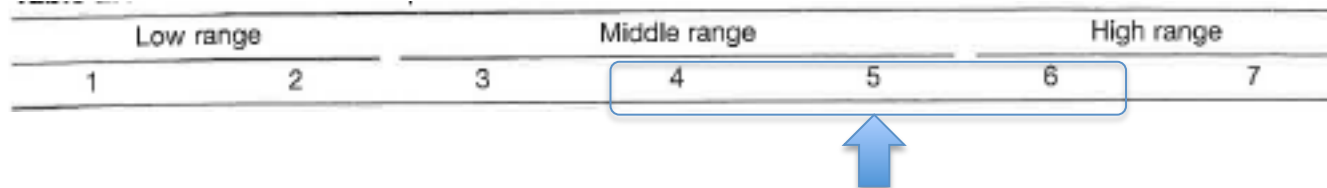
# CLASS Inter-Rater Reliability: Is “Within One” Good Enough?

The CLASS (like the ECERS-R and other observational systems) assesses agreement “within one” encompassing broad regions of the 7-point scale.

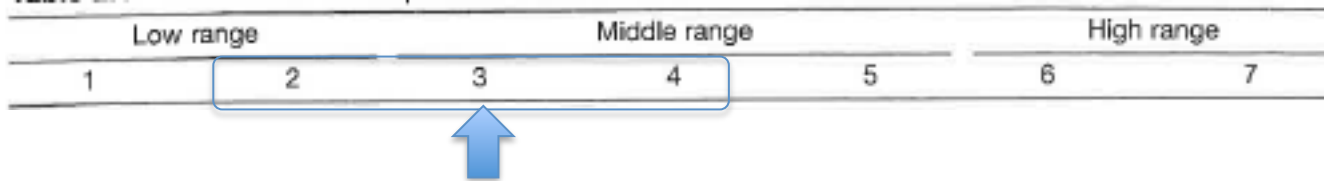
A score of 5, 6 or 7 is considered in agreement with a master score of 6.



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.



## Exact Agreement Difficult on the Highly Inferential System

**Table 2.1.** Dimension descriptions for the CLASS

Low range		Middle range			High range	
1	2	3	4	5	6	7
The low-range description fits the classroom and/or teacher very well. All, or almost all, relevant indicators in the low range are present.	The low-range description mostly fits the classroom and/or teacher, but there are one or two indicators that are in the middle range.	The middle-range description mostly fits the classroom and/or teacher, but there are one or two indicators in the low range.	The middle-range description fits the classroom and/or teacher very well. All, or almost all, relevant indicators in the middle range are present.	The middle-range description mostly fits the classroom and/or teacher, but there are one or two indicators in the high range.	The high-range description mostly fits the classroom and/or teacher, but there are one or two indicators in the middle range.	The high-range description fits the classroom and/or teacher very well. All, or almost all, relevant indicators in the high range are present.

## 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 and a big 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 Feedback<sup>9</sup>

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

### Scaffolding

- Hints
- Assistance

### Feedback loops

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

### Prompting thought processes

- Asks students to explain thinking
- Queries responses and actions

### Providing information

- Expansion
- Clarification
- Specific feedback

### Encouragement and affirmation

- Recognition
- Reinforcement
- Student persistence

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

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

The teacher gives only perfunctory feedback to students.

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

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

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

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

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

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

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

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

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

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

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

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

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

QUALITY  
OF FEEDBACK

<sup>9</sup>Quality of Feedback is generally observed in response to a student's or students' answer to a question or as a student progresses on his or her work or involvement in an activity, whereas Concept Development is the method a teacher uses as he or she provides instruction or activities.

# Challenge of Rater Variance

- Based on Head Start training, CLASS developers (Cash, Hamre, Pianta, & Myers, 2012) reported:
  - Exact agreement was low: 41% overall exact agreement with master score in training of over 2,000 Head Start staff.
  - Black and Latino raters placed their Instructional Support scores farther from the master score as did raters who disagreed with intentional teaching beliefs.
- Recent report on rater errors in CLASS-S (McCaffrey et al., *Educational Measurement*).

# Conclusions

# Summary: Limits of Adopting Existing Measures for High Stakes Use

- When scrutinizing these measures which were *developed for other purposes*, it is not surprising that there are *limitations for the ways in which they have been adopted for high stakes policy uses*.
- The limitations of the reliability and validity evidence for ECERS-R and CLASS may, in part, explain the relatively low associations with children's developmental gains during preschool.

# Alternative Approach to Evaluating Level of Evidence

- Consistent with the latest *Standards for Educational and Psychological Testing* may need to step back and 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 short

- A measure is not statically “reliable and valid”
- The evidence must be fully evaluated and regularly revisited (including locally) for each use.
  - The body of evidence needed to demonstrate reliability and validity for program *self-assessment*
  - May be different from reliability and validity for teacher *professional development*
  - Which may be different from reliability and validity for *policy decision making and accountability*

# Alternative Approach to Evaluating Level of Evidence

- As a concrete example, 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 need agreement on definitions of readiness and the ***aspects of quality*** that support them, and measures designed and evaluated to assess those aspects of quality.

# Consider Alternative Approaches to Accumulating Evidence

- Continuous and local validation and improvement of measures.
- This approach could potentially 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).

# Local Validation Can Also Encompass Contextual Diversity

- Does a measure capture a single conception of quality?
  - Is that conception explicit or implicit?
  - Does that conception match with policy goals and with on-the-ground practice, across local contexts?
- Does a measure capture well all children's experiences?
  - Or, “average” teacher quality, “average” child, “substantial portion of the day”

# Future Directions

- New ECERS-3
  - Some important changes (e.g., expanded content on language, literacy and math).
  - Others remained the same (encourages scoring all indicators, and alternative scoring in development, but manual retains stop scoring).
  - IES-funded validation study of ECERS-3.
- New measure development
  - IES Research Network on Early Childhood Education.
  - Includes assessment team (headed by Carol Connor).
  - New measure development for QRIS.

# Future Directions (cont)

- Early Investments Initiative (Gordon, Zinsser, Sheridan, Main, Curby, et al., **IGPA**) & EMOTERS (Zinsser, Curby, Gordon, et al. **IES R305A160010**)
  - New measure of Social and Emotional Teaching.
    - The variety of activities and practices that promote SEL in children.
  - One of the novel design features: video weeks
    - Video full week per classroom (panoramic & closeup).
    - Easier to have many coders, to examine IRR.
    - Facilitate identification of within/across day variation.
    - Using in iterative measure development.
    - Using generalizability theory to parse sources of variation.

# Future Directions (cont)

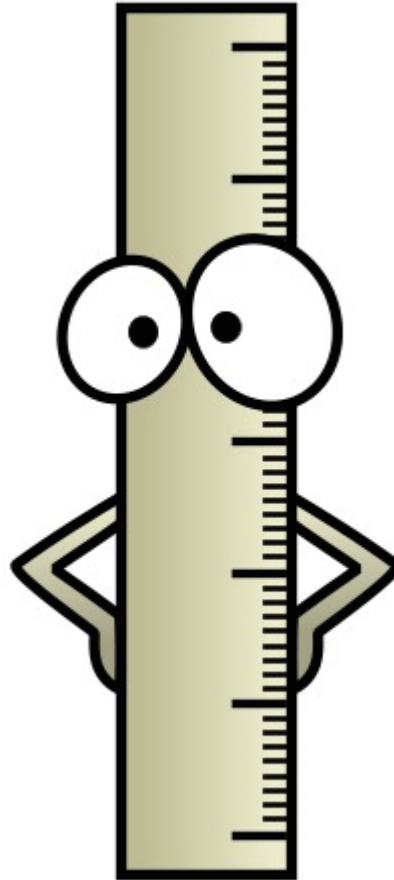
- **IES R305H130012**
  - Researcher-Practitioner Partnerships in Education Research grant
  - Creating a Monitoring System for School Districts to Promote Academic, Social, and Emotional Learning: A Researcher-Practitioner Partnership
  - CASEL & Washoe County School District
- Using IRT (Rasch) approach during iterative item development.
  - Refining construct definition
  - Developing item pool
  - Continuous refinement
  - Anchor (common) and new (unique) items

# The Rasch Ruler

Measures =  
Kids' Levels

Kids who have the MOST  
competency

Kids who have the LEAST  
competency

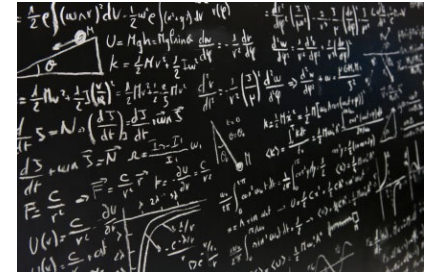
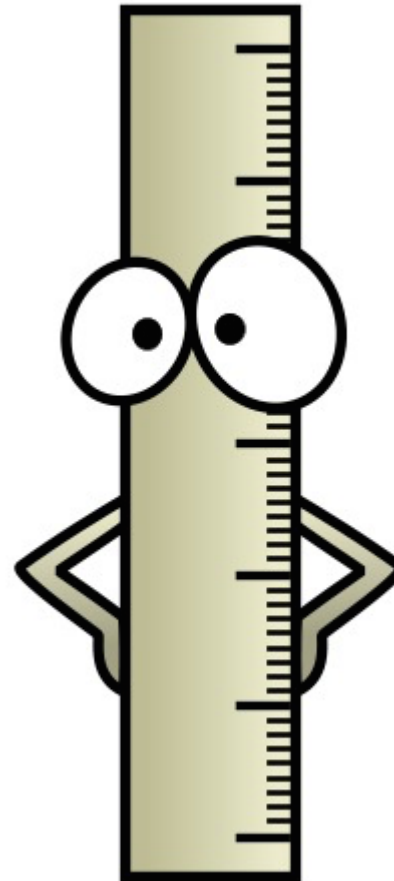


Marks =  
Competencies

Competencies that are  
really HARD for most  
kids

Competencies that are  
really EASY for most kids

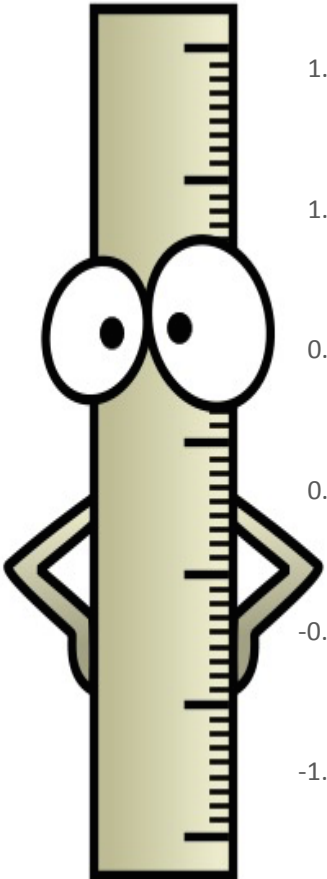
# The Rasch Ruler



**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 our most improved 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.

5. Getting along with my classmates.

1. Sharing what I am feeling with others.

4. Introducing myself to a new student at school.

6. Being polite to adults.

