Challenges Aligning Existing Measures with Professional Development Learning Goals to Evaluate Program Effectiveness

Nicole B. Kersting, University of Arizona

Amy Brodesky, Education Development Center

Jim Hammerman
STEM Education Evaluation Center (SEEC) at TERC

Common Challenge

The tension in selecting or designing measures between:

- Alignment with a program's learning goals to better assess treatment effects (often by creating customized measures)
- Comparability of research results across studies to build broader field knowledge (often by using existing measures with proven properties)

Session Goals

- Provide a forum for discussion of issues and tensions in selecting measures to assess program effectiveness that align with learning goals
- Share experiences and strategies for addressing these issues

Session Overview

1. Aligning PD Program Goals and Measures

   Discussion

2. Describing Some of the Tensions and Trade-offs

   Discussion

3. Issues in Creating Measures

   Discussion

4. Group Brainstorm: Considerations and Strategies for Addressing these Issues

Topic 1: Aligning PD Program Goals and Measures

- An example from one PD program
- Discussion

Differentiated Professional Development: Building Mathematics Knowledge for Teaching Struggling Learners

- **PD Programs** on Common Core State Standards for: 1)Fractions 2)Decimals 3)Positive & Negative Numbers
- **Audience:** General ed./math teachers & special educators (grades 4-7)
- **Blended Format:** 2 face-to-face workshops & 6 session online course (14 weeks long, 48 PD hours)

Differentiated PD Model

1) Core Activities that everyone does

- Essential content
- Common experiences

2) Choice Points that allow teachers to choose options and activities to individualize their learning

3) Self-Assessment and Reflection Opportunities to help teachers reflect on their own understandings and guide their choices

PD Goals for Teachers: *Fractions Example*

- **Math Content:** Build a deep and flexible understanding of fraction concepts (CCSS grades 4-6)

- **Diagnostic Approaches**: Build knowledge and skills for identifying students' strengths, difficulties, and misconceptions

- **Instructional Practices**: Expand repertoire of research-recommended practices for teaching fractions to struggling math learners

Challenges

- How to measure impact of the PD program on teachers' knowledge and practices
- Tensions between creating a PD program that is innovative and that aligns well with existing research measures

Our Considerations in Choosing Measures

- Alignment of measures with PD goals
- Lessons learned from prior projects
- Implementation Concerns
  - Will our audience of general and special educators feel comfortable taking the instrument?
- Time and $$ for data collection & analysis
- Overall set of measures: mix of established and project created

Why did we add the Classroom Video Assessment (CVA)?

Alignment with goals

- Content: 13 fraction videos available
- Rubrics: Math Content, Suggestions for Improvement, Student Thinking, and Depth of Interpretation

Other Considerations

- Online data collection
- Established rubrics and trained scorers
- Possibility of computer scoring

How does CVA fit with the Set of Measures?

Pre-post measures

- Fractions math assessment (MKT)
- **Classroom video assessment (CVA)**
- Fractions pedagogy survey

Other instruments

- Background survey
- Post-program evaluation survey
- Post-program interview
- Online course data
- Classroom observations for case study teachers

Alignment: PD Program & CVA

- Areas of close alignment
  - Fractions content
  - Student thinking
- Some PD areas are not specifically addressed by measure
  - Diagnostic probes
  - Struggling learners
  - Specific representations

Discuss in Small Groups

Think about the measures you included in a proposal for a current project.

- How well do those measures align with your project now?
- If you added or changed a measure, what were your reasons for doing so?

Time: ~10 minutes for small group discussion followed by whole group share-out

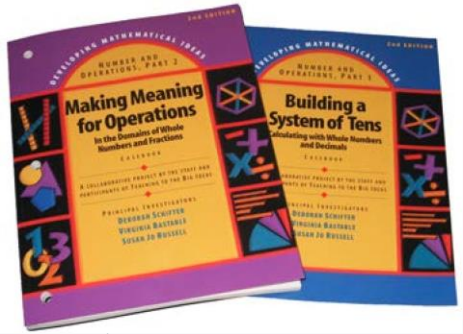
Topic 2: Describing Some of the Tensions and Trade-offs

- Examples from the Evaluating DMI PD program
- Discussion

Navigating Tensions and Trade-offs in Measure Selection and Creation in *Evaluating DMI*

Jim Hammerman

STEM Education Evaluation Center (SEEC)
TERC, Cambridge, MA



What is *Evaluating DMI*?

- Experimental study of the impact of the *Developing Mathematical Ideas (DMI)* PD on elementary grades (1-5) teachers
- PD focus on understanding math (number and operations content) and student thinking, and teaching to support mathematical reasoning
- Impact on teacher knowledge, teaching practice, and student achievement, & their relationships
- Also: PD Logs, reflective survey, interviews

Multiple Measures of Each Construct

- Teacher Knowledge
  - LMT Math Knowledge for Teaching (MKT)
  - Open-Ended DMI-aligned measure
  - New CVA focused on number and operations
- Student Achievement
  - Curriculum aligned end-of-unit and end-of-year assessments from Boston Public Schools
  - State- or District-wide assessment as available

Tensions to Navigate: Teacher Knowledge

- MKT: Widely used, easy to use, multiple forms (pre-, post1-, post2-), related knowledge, *but* not everything we care about
- Open-ended DMI: Well-aligned, prior instrument development and testing work, *but* harder to score, not widely used
- CVA: Knowledge grounded in practice of theoretical interest, structure and system in place, *but* need to create particular content

Tensions to Navigate: Student Achievement

- State- and District-wide tests: Wide use, of policy interest, content partially related, *but* not all grade levels, differs by state, doesn't measure key intended impacts
- BPS EOU and EOY assessments: Better alignment, can be used in all districts, *but* not equated across grades, need to negotiate use outside BPS, still doesn't measure key impacts
- Other issues: Burden, cost, access

Teaching Practice Measures

- Existing scales
  - Reform Teaching Observation Protocol (RTOP)
  - Instructional Quality Measure (IQM)
  - Mathematical Quality of Instruction (MQI)
    - Original and revised scales
- Invented scales
  - New measure built on Stein & Kaufman curriculum implementation theory
  - New student engagement rubrics, based on work by Webb, Franke et al.

Tensions to Navigate: Teaching Practice

- RTOP: Existing measure, wide use, easy to use, *but* doesn't capture key constructs of interest
- IQM: Relevant constructs, some external use, *but* practical and cost constraints
- Stein & Kaufman measures: Important constructs, *but* hard to implement reliably
- MQI: Existing measure, wide use, training & calibration systems, important content, *but*…
- Student Engagement rubrics: …to address missing key content, *but* reliability issues

Discuss in Small Groups

- How do the issues identified here resonate with issues you've addressed in your projects?
- What other tensions have you grappled with?

Topic 3: Creating Measures

- Important Issues and Challenges for Instrument Development
- Two examples: The Classroom Video Analysis (CVA) instrument & the Instructional Quality in Mathematics Instrument
- Discussion

The Classroom Video Analysis (CVA) Instrument

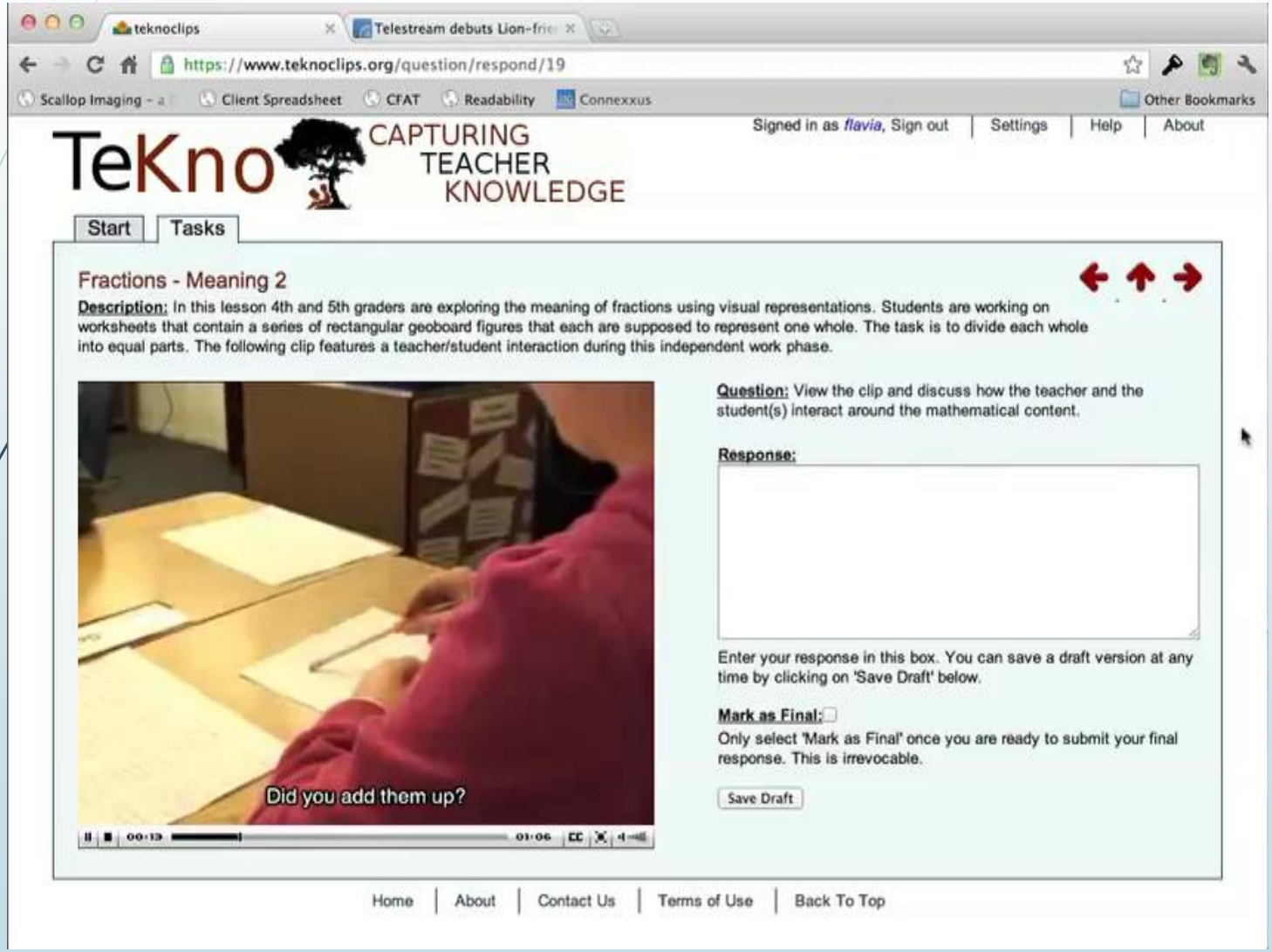
What do we want to measure?

- Designed to measure usable teaching knowledge in mathematics
- Defined as the knowledge teachers can access and use during instruction

How do we measure it?

- Uses authentic video clips of classroom instruction as item prompts to elicit usable knowledge
- Based on the analysis of teaching as important teaching skill
- Scoring of written responses provides measures of usable teaching knowledge

Let's take a Look .... https://www.teknoclips.org



Instructional Quality in Mathematics Instrument

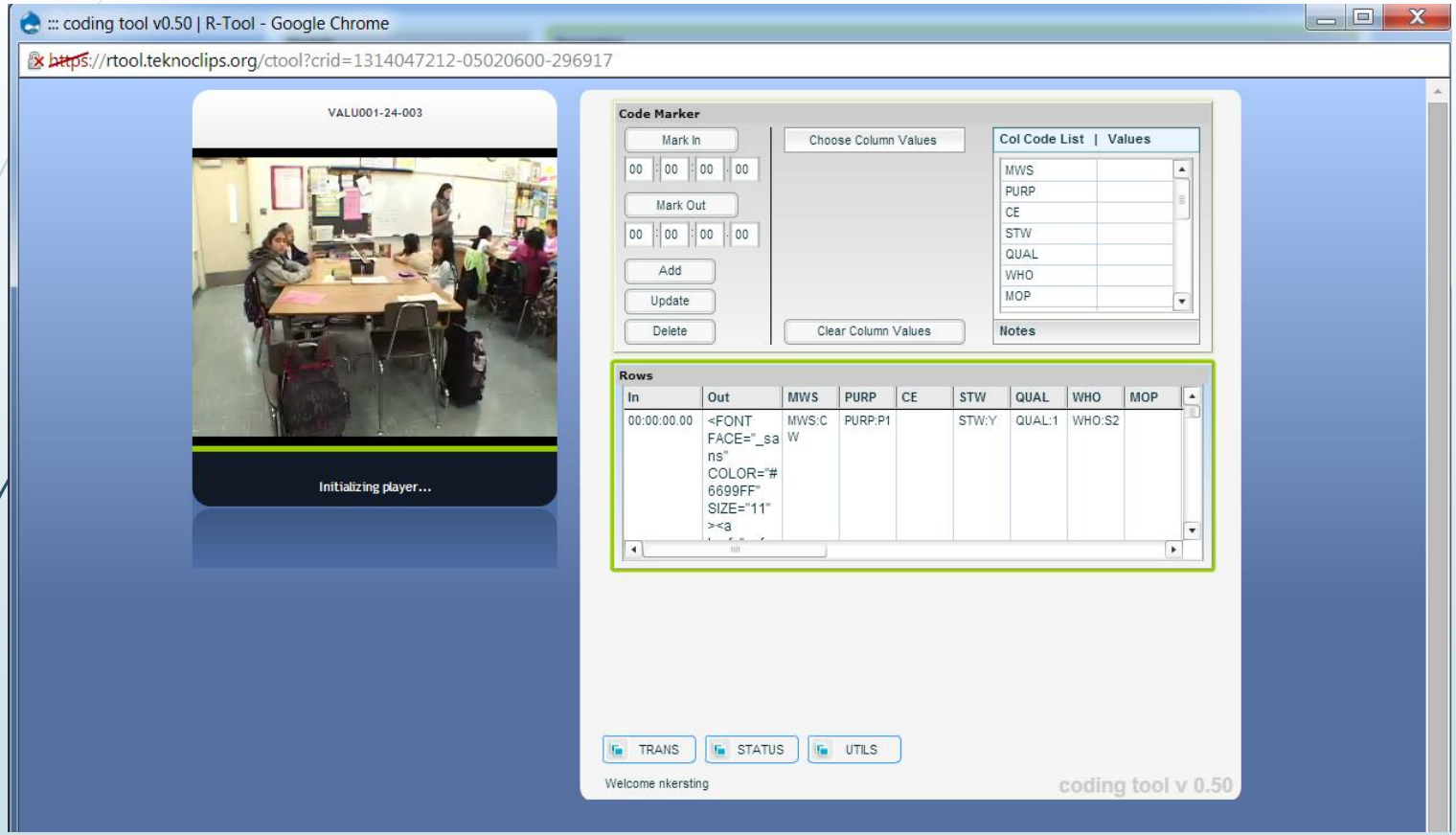
What?

- Mathematics Quality of Instruction
- What kind and amount of mathematical work do students do & are mathematical ideas and concepts made visible (Hiebert & Grouws, 2007)

How?

- Rubrics that address aspects of the two larger themes
- Ordered Categories
- Proportion of time spent on high quality instructional features represents overall instructional quality score

Taking a look



What we get out …

VideoCaseTitle	case	Prop High QualityInteraction	Prop Modeling	Prop Time New	Prop Seatwork	Prop Struggle	Prop Mapping Algorithms	Prop Concept Developed	Total IQ
VALU001-01-001	1	0.03	0.13	0.42	0.32	0.00		0.34	1.25
VALU001-01-002	1	0.02	0.49	0.07	0.50	0.00	0.11	0.00	1.19
VALU001-01-003	1	0.43	0.66	0.14	0.51	0.41	0.00	0.00	2.15
VALU001-01-004	1	0.19	0.28	0.52	0.90	0.73	0.42	0.18	3.22
VALU001-01-005	1	0.16	0.15	0.09	0.45	0.00	0.00	0.05	0.89
VALU001-02-001	2	0.33	0.51	1.00	0.86	0.00		0.00	2.69
VALU001-02-002	2	0.47	0.76	0.72	0.63	0.00		0.32	2.90
VALU001-02-003	2	0.31	0.47	0.30	0.59	0.39	0.00	0.00	2.06
VALU001-02-004	2	0.62	0.28	0.82	0.87	0.18	0.00	0.00	2.77
VALU001-02-005	2	0.48	0.31	0.27	0.66	0.00	0.10	0.00	1.82
VALU001-03-001	3	0.15	0.73	1.00	1.00	0.00		0.21	3.09
VALU001-03-002	3	0.08	0.62	0.28	0.75	0.00		0.08	1.82
VALU001-03-003	3	0.27	0.47	0.00	0.74	0.00			1.48
VALU001-03-004	3	0.31	0.45	0.26	0.73	0.33	0.00	0.00	2.09
VALU001-03-005	3	0.14	0.15	0.49	0.06	0.00	0.00	0.00	0.83
VALU001-04-001	4	0.60	0.50	1.00	0.92	0.92		0.58	4.52
VALU001-04-002	4	0.46	0.94	0.52	0.64	0.71		0.00	3.28
VALU001-04-003	4	0.57	0.62	0.25	0.73	0.80		0.21	3.17
VALU001-04-004	4	0.65	0.57	0.28	0.96	0.60	0.00	0.12	3.17

Creating Measures

- Naming & Defining the construct (What?)
- Operationalizing the construct (How?)
- Issue and Challenges:
  - Identifying a relevant or adequate construct of latent trait
  - Often no well developed theory
  - Grain size issue
  - Good Alignment versus "overfit"

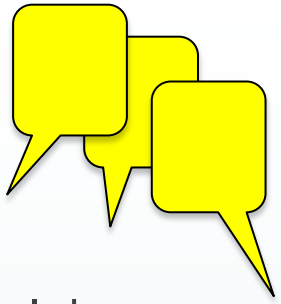
Issues and Challenges Continued

- Identifying what will not be measured
- Item design
- Items as indicators of construct
- Authenticity
- Combining subscores
- Sensitivity of measures to assess change

Scalability of Measures

- Project use only versus wider use in the field
- Issues of score reliability across raters and over time (ongoing reliability information as opposed to initial and midpoint)
- CVA: human raters score teacher responses according to four rubrics
- IQM: human raters score lesson videos
- Standardization / Automation using machine learning approaches
- Feasibility & Cost and time effectiveness

Discussion Questions



(1) Which issues and challenges are you or would you be facing in your project if you decided to develop project specific measures?

(2) What kind of decisions have you made regarding some of these issues and challenges and what have you learned from the decisions you made?

(3) What kind of trade-offs between existing and project developed measures would you face in your project?

Time: ~10 minutes for small group discussion followed by whole group share-out

Key tensions …

- **Most important issues generated from the final participant discussion**

- New versus used
- Legitimacy of used versus new
- Feedback to teachers – Who is the audience?
- Risk of instrument being used for different purpose than it was designed (i.e., potential mis-matches between original instrument design and current intended use -- purpose, population, time-scale)

Key tensions continued …

- Alignment – too close to treatment
- Grain size
- Cost & Logistical & Expertise
- Multiple teacher observations with different instruments challenge for teachers (district or state mandated accountability measures versus research project measures)
- Different quality observational measures
- How do research instruments fit in with measures already in use
- How do we get teachers to care about faithfully completing measures
- How do we invite research participants to get buy-in for the research component of the study?

How do we navigate these tensions?

- **Things to consider generated by the final participant discussion….**
- Consider original purpose of the instrument carefully
- Consider population for which instrument was designed
- If you plan to develop instruments team up with a psychometrician
- Seeing an instrument development effort all the way through
- Do an initial test of the instrument versus scaling up

How to navigate these tensions continued ….

- Identify important resources such as bank of measures
- Consider measures used for state funded MSP grants
- Sharing measures in development to make a more systematic effort toward measure development
- Using new measures in conjunction with existing measures
- Identifying things that aren't being measured
- Create a database of measures that has important design and use information (including informal feedback from Pis who have used it in their studies) easily accessible
- Mary Kennedy – effects of different kinds of measures