Accountable assessment

Professor Richard Lehrer

Vanderbilt University, Tennessee

https://doi.org/10.37517/978-1-74286-638-3_9

Richard Lehrer is Professor Emeritus and Research Professor of Education at Vanderbilt University. A former high school science teacher, he is a member of the National Academy of Education, a Fellow of the American Educational Research Association and a recipient of the American Psychological Association's Distinguished Contributions in Applications of Psychology to Education. Working with teachers, he focuses on the design of classroom learning environments that support the growth and development of children's understandings of how knowledge is generated and revised in mathematics and in sciences. His research examines productive means to engage children in forms of mathematics that are critical to STEM education, especially geometry, measure, data and chance. This research is aligned with parallel efforts in science education to induct children in the signature practice of the sciences—invention and revisions of models of natural systems. A contemporary collaboration with the Berkeley Evaluation and Assessment Research Center explores the feasibility of integrating teachers' classroom-embedded judgements of student learning into the kinds of psychometric models that are employed in standardised, 'accountability' assessments.

Abstract

There is widespread agreement about the importance of accounting for the extent to which educational systems advance student learning. Yet, the forms and formats of accountable assessments often ill serve students and teachers; the summative judgements of student performance that are typically employed to indicate proficiencies on benchmarks of student learning commonly fail to capture student performance in ways that are specific and actionable for teachers. Timing is another key barrier to the utility of summative assessment. In the US, summative evaluations occur at the end of the school year and may serve future students, but do not help teachers to better support the students who were tested. In contrast, formative assessments provide actionable grounds to improve the quality of instruction on the basis of both the granularity and specificity of their content and their timing. Unfortunately, the psychometric qualities of formative assessments are often unknown. I describe an innovative approach to assessment that aims to blend the productive characteristics of both summative and formative assessment. The resulting assessment system is accountable to students and teachers by providing actionable information for improving classroom instruction, and at the same time, it addresses the demands of psychometric quality for purposes of system accountability as it is currently practiced (in the US). The innovative assessment system relies on partnership with teachers to generate 1) a shared conceptual frame for describing instructional goals and valued forms of teaching and learning; 2) a set of electronic tools to help teachers detect, share, analyse, and interpret student learning data; and 3) classroom and school-level community professional development structures to support and sustain a widespread practice of assessing to guide instruction. These features are coupled with new psychometric models, developed by the Berkeley Evaluation and Assessment Research Center, that provide more robust estimates of student learning by linking information from multiple sources, including student classroom work, student responses to formative assessments, and summative evaluations. (Mark Wilson will address the psychometric modeling during this conference.) Here I describe challenges

and prospects for this innovation with a case study of its implementation in a K–5 elementary school that is seeking to improve the quality of instruction and students’ understandings of measure and rational number arithmetic.

Introduction

Although the purposes of assessment are varied, there is widespread agreement about the importance of accounting for the extent to which educational systems advance student learning. Yet, the forms and formats of accountable assessments often ill serve students and teachers. In the US, summative evaluations used for accountability occur at the end of the school year. These evaluations could, in principle, serve future students, but they do not help teachers better support the students who were tested. Moreover, the implications of student performance on these summative evaluations for instruction tend to be very general, primarily because the tests are constructed in ways that are not well informed by constructs that describe typical progressions and patterns of student thinking (Wilson, 2005). As a result, knowing that student performance in any area of mathematics is substandard does little to inform specific steps toward instructional improvement. In contrast, formative assessments are designed to provide actionable grounds to improve the quality of instruction due to increased granularity and specificity of their content and their timing (e.g. Black & Wiliam, 1998, 2009). As Wiliam (2015) clarifies, the signature of formative assessment is anticipating how students will think about situations posed during assessment and taking appropriate action accordingly. Unfortunately, the psychometric qualities of these forms of assessment are often unknown, and therefore are difficult to align with accountability assessments.

The premise of our collaboration with colleagues at University of California, Berkeley is that if ongoing assessment of student thinking is woven into the fabric of instruction, then teacher judgements of students’ ways of thinking can inform psychometric modelling of student learning. Summative and ongoing formative assessments can be coordinated to generate more robust and actionable accounts of student learning. Moreover, assessment can be more accountable to the ongoing improvement of instructional practice and student learning in real time, rather than serving primarily as an aftermath to instruction. Achieving these goals means that teachers must learn to read and register selected forms of student thinking as they emerge during the course of classroom activity. Moreover, on the basis of what the data show, teachers must learn to leverage their knowledge of student thinking to improve the quality of instruction, so that assessment becomes a vital part of instructional practice.

Moreover, although most assessments are conducted by individual teachers, the practice of assessment, as well as its meaning and perceived value, are influenced by the surrounding community (Horn et al., 2015). In workgroups and grade teams, teachers communicate and subtly enforce a common epistemic orientation toward assessment (Horn et al., 2015). By epistemic orientation, Horn means teachers’ perspectives – often tacit – on what can be known with data, how to know it, and why it is of value. Consequently, assessment practice is constituted by an interplay between individual teacher activity in a classroom or related instructional setting, and a teacher’s anticipations of the norms and interpretations of the surrounding community. With this dual view of assessment practice in mind, we aimed to create and test an assessment system designed to address two coordinated purposes: 1) to provide ongoing, instructionally-productive evidence to teachers about student learning; and 2) to link dense information from student work products and formative assessments with summative assessments in new psychometric models that generate robust estimates of the growth of student learning. Such an assessment system includes:

- a conceptual frame shared by all participants for generating and interpreting evidence of learning in student activity across instructional settings

- a set of tools to amplify teachers' ability to detect, capture, share, analyse, and make sense of evidence of learning across instructional settings
- community structures across classroom, school and project partnership to support and sustain the practice of assessing to guide instruction.

To test the feasibility of this innovation, we collaborated with K–5 teachers in an intact school setting to construct an assessment system that would allow us, collectively, to track student learning of the mathematics of measure (length, angle, area, volume), and of children's learning of related concepts of rational number as teachers introduced measurement models to promote learning about fractional quantities and operators. The initial impetus for the focus on measure was children's comparatively poor prior performance on summative, statewide assessment in these areas of mathematics, as well as its many conceptual connections to a wide array of mathematical concepts taught in the elementary grades.

Constructing an assessment system

Participants

To construct the elements of an assessment system – a shared conceptual frame, appropriate tools, and productive community structures – we collaborated with 18 K–5 teachers, most of whom taught at Sleeve Elementary in the south-central region of the US. Three participating teachers were located at another school in the district. The district is the largest in the state. The student population of Sleeve Elementary is primarily rural and white. I met with teachers once each month for two to three days over two years (Summer 2018 – May, 2019; September 2019 – March 2020, interrupted by the suspension of schooling due to the COVID-19 pandemic). I also conducted multi-day summer institutes each year, once in person and once via Zoom conferencing. During the past year (August 2020– present), students attended school in person intermittently, and instruction was conducted online during the rest of the time. Access to digital instruction was especially problematic for many students.

Conceptual tools to promote shared vision

Supporting teachers to articulate a shared vision of instruction, learning and assessment included the design and iterative development (with teacher feedback and frequent contribution) of a set of conceptual tools. These included most prominently *constructs*, *lessons* and formative *assessment items* to support student learning of particular elements of constructs.

Constructs

Constructs identify typical forms of student thinking and articulate how these forms of thinking progress when they are appropriately supported by instruction (Wilson, 2005). The constructs are not fully-fledged theories of learning, but rather, are tuned to highlight aspects of learning that contribute to effective next instructional steps within specific content areas. Theories of learning are necessarily much finer-grained and more technical, and are not usually accurately described as linear (stage-like) paths through levels of a construct (Lehrer & Schauble, 2015).

Progress maps describe how children's thinking, as captured in constructs, usually develops. Progress maps are coarser-grained descriptions that are intelligible and practical; they represent an informational tradeoff for informing instruction. That is, they capture important variants in student

thinking, but like all models, omit variations less commonly observed and forms of thinking that are not usually useful for guiding instruction. They set a local mathematical horizon that influences how teachers respond to students during the course of formative assessment. That is, they help teachers identify local ‘next steps’ in student thinking, so that they can decide upon reasonable approaches for supporting students’ learning without having to manage a level of information that would otherwise be overwhelming (Kim & Lehrer, 2015).

We developed and refined four constructs that depict student progress in conceptions of the measure of length, angle, area, and volume. The constructs are organised as narratives of development and are summarised as tables of levels that describe and exemplify growth in students’ ways of thinking. Each level is constituted by multiple sub-levels that collectively constitute the form of thinking characterised by that level. For example, initial levels (Levels 1 and 2) of the length construct specify how young children first begin to engage with the fundamental problematic of measure – identifying and characterising attributes to be measured and comparing values of these attributes directly and also indirectly via units of measure. Performances at these initial levels focus on properties of unit, such as the need to tile units without gaps or overlaps, and on understandings of the logical necessity that governs the performance (e.g. why gaps or overlaps produce inaccurate measures, not simply that they do). Figure 1 lists the levels of the length construct and, for Theory of Measure – Length (ToML) Level 2, illustrates how each level is composed of a network of related concepts that collectively are indexed by that level.

Figure 1 Theory of Measure – Length (ToML)

Lessons

Classroom lessons are designed to clarify how the conceptual change envisioned by constructs can be supported by instructional practices. For example, an image of length as dynamically generated by travelling from a starting point to a specified location often helps young children conceptualise length as a distance. This interpretation makes symbolisation of units on a ruler more intelligible, so that the location of 1 at an endpoint of a unit interval is interpreted as the distance travelled, rather than as merely marking one unit of a collection of units. Over the course of our collaboration, the lessons have undergone multiple rounds of revision and have been augmented with teacher-authored examples and alternatives represented in a ‘teacher’s corner’. Teachers and researchers regard lessons not as static structures, but subject to change as we collectively learn more about student thinking and how to support it.

Formative assessment

Every lesson includes formative assessment items and illustrates how prospective student responses are aligned with particular levels on the construct map. For example, one of the formative assessment item displays six two-dimensional figures (including a line and a figure that is not closed) and asks students to circle all the figures that have an area. After students complete the formative assessment, and after the teacher has aligned student responses to levels of the relevant construct (in this case, the area construct), the teacher conducts a *formative assessment conversation* in which they juxtapose student responses and students explain the thinking that guided their responses.

In a follow-up discussion about the item just described, some Grade 3 students (7–8 years old) argued that it is possible to find an area measure for figures that are ‘almost’ closed. Rather than rejecting this proposal, the teacher asked children to justify their choices. At the board, students demonstrated how they would tile the entire space into which the area ‘leaked’. Other students agreed that they could obtain a measure in this way, but objected that it would be difficult to know when to stop. Should one ‘go to the road’ outside the school? The teacher then drew ‘large’ and ‘small’ open figures, asking children to estimate the area measurement of each. Children concluded that all open figures would have the same (infinite) measure and conceded that this result would defeat the original intent to use measure to compare areas. Thus, rather than resorting to pre-determined definitions, the teacher supported students in reaching the consensus that it made most sense to restrict area measure to closed figures.

Constructs, lessons (including teacher elaborations), and formative assessments are available digitally, as illustrated in Figure 2. The district has adopted many of the lessons to guide their mathematics education program, although that also has had the unfortunate consequence that lessons have been incorporated into pacing guides and related forms of curricular control.

Figure 2 A suite of conceptual tools: lessons, constructs and formative assessments

THEORY OF MEASUREMENT - LENGTH

Length Unit 1

Measuring Is Comparing

[Complete Lesson](#) [Formative Assessment](#)

MEASURING IS COMPARING

Measurement is a way of comparing the magnitude (extent) of an attribute from one object to the next. For example, one person is taller than another person. This unit introduces students to the idea of the extent of comparison by directly comparing the magnitude of the length and circumference features of different objects.

Mathematical Concepts

- Measurement is the act of comparing the magnitude (extent) of the same attribute for two objects (or different objects).
- An attribute is a characteristic of an object, such as its length, weight, height, or width.
- Every object can be described by multiple attributes.
- Attributes used to describe an object are measured and compared in the same way (e.g., length is measured by using the same unit).
- Some attributes, such as length, have continuous (uninterrupted) scales. Other attributes are primarily discrete, such as volume and color.
- The magnitude of the same attribute of two different objects can be compared directly either by physical comparison or by comparing measurements of them.
- Objects can be sorted by the magnitude of one or more attributes.

Length Unit 1

Children first look at 3-4 pumpkins and tell what they notice about them. They typically notice attributes such as texture ("smooth," "bumpy"), size ("big," "little"), circumference ("fat around"), color, weight ("heavy," "light") and height ("tall," "short"). What children notice helps establish a big idea: a pumpkin, like any other object, can be described by multiple attributes. The teacher asks which of these attributes or qualities will help determine the "biggest" pumpkin. Different senses of "bigness" are elicited from children.

Teacher's Corner

[pdf Julie Measure - Up: Describing Plant Growth in Kindergarten](#)
We used the context of plant growth to introduce scientific practices of observing, representing, and measuring. Children worked in teams to plant Amaryllis bulbs in pots, one bulb per pot.

[pdf Rachel Comparing Weights with a Pan Balance](#)
Students order 4 objects by weight with a pan balance. It is more interesting if one of the objects has a greater volume but it in fact weighs less. For example, a can of fruit and a Styrofoam box.

[Submit a Teacher's Corner entry](#)

Performances

ToML1A: Pose a question or make statements about a potentially measurable object of interest.

ToML1B: Identify measurable attributes (qualities).

Digital tools to support ongoing assessment (designed by Corey Brady)

Teacher observation tools (TOTs)

Teachers' judgements of students' ways of thinking are recorded with a web-based toolkit implemented on iPads. The toolkit allows teachers to record and store evidence of student thinking (typically video, photo, and field notes) that they observe during the course of instruction, and to associate this evidence with particular sub-levels of one or more constructs by means of a built-in coding system. This capacity extends the meaning of 'item' to include diverse expressions of student thinking as revealed by student talk, activity, and work products. Figure 3 is a facsimile of the recording portion of the toolkit. It exemplifies a photo and teacher note, the teacher's selection of the appropriate construct sub-level that describes one or more students' thinking, and attribution to one or more students.

Figure 3 Recording evidence of student thinking

Teacher Observation Tools

Add photo or video:

Add notes:

"when you have spaces and gaps, the measurement is a smaller number because the marker isn't covering that part."

Add Students:

S1 S2 S3 S4

Assign construct level(s):

ToML 2A ToML 2B ToML 2C

Learning construct details:

ToML 2B | Tile and explain why (the explanation is required)

Learning construct examples:

"The units should touch so there is no gap. Gaps mean that some of the space is not measured."
"The units should not overlap. If they do, you measure some of the space twice."

TOTs includes visualisations of student data that serve several functions – some for individual classroom teachers and others at a community-wide level. Figure 4 displays a facsimile of a dot plot of evidence for a construct from one teacher's classroom. Each dot corresponds to an observation and when selected the contents of the observation are revealed (here a portion of the previous observation is displayed). This display is handy for tracking evidence at the construct level for the class and provides a general picture of the class's current progress with respect to the given construct.

Figure 4 Dot plot of observations by construct sub-level.

A more economical display of data like these that seems to be preferred by teachers is a ‘heat map’ (see Figure 5), which uses color intensity represent frequency of observation. This view can also be used to represent observations across classrooms. This school-wide view is an important component of an emerging assessment practice in the school that is described in the next section.

Figure 5 Heat map of observations by construct sub-level across classrooms.

A ‘star chart’ view, depicted in Figure 6, represents observations at particular sub-levels of a construct for individual students, a feature that helps teachers ensure that their estimates of student learning are based on a census of students, and not a select few.

Figure 6 Student level evidence of learning

Establishing a community of assessment practice

We collaborated with teachers to establish practices of assessment that were supported by the conceptual tools of constructs, lessons and formative assessments, and by the use of TOTs to generate evidence of student learning. As noted previously, our emphasis on community was informed by its critical role in the development of the professional discourse necessary for the improvement of instruction (e.g. Ball & Cohen, 1999, Desimone, 2009; Gibbons & Cobb, 2017) and by its critical role in generating productive norms for assessment (Horn et al., 2015). We faced several challenges in realising a collective vision. For many teachers, these forms of mathematics were not familiar, primarily because past instructional practice in the school had emphasised procedural competence with tools, such as protractors and rulers. A related challenge was that instructional practices did not include a repertoire of ways of helping students conceptualise measure. Instead, the sole focus was on whether a measure proposed by a student was or was not correct. Other challenges included the nature of the conceptual tools available to teachers. Initially, we represented constructs describing the progression of student thinking as tables. These brief descriptions had the virtue of economy but they did not communicate well. Similarly, our initial attempts at lessons were not sufficiently educative – they did not reveal why particular tasks and tools were likely to support student learning. And at first the observation tools were in embryonic form. However, teachers already had a history of exploring the growth of student thinking in other realms of mathematics, especially whole-number arithmetic. As a consequence, our efforts to develop a community of practice centered around student thinking was well received. In this light, we engaged in several forms of community building.

Learning labs

We adapted ‘math labs’ (Kazemi et al., 2018) to collaboratively generate opportunities to learn from and with students. During a learning lab, teachers collaborated to plan, conduct and reflect upon student learning in situ. Teachers were sometimes grouped by grade band (e.g. K–2, 3–5) and at other times constituted across grades (K–5). An instructional facilitator and I assisted at every lab (two or three labs per day were conducted at each of my monthly visits). The initial phase of the lab

included decisions about a portion of a lesson that would serve as a focus. The group anticipated how students might think about this portion – in the language of a construct – and what we were especially interested in seeing in more depth. Occasionally teachers reviewed the mathematical concepts beforehand so that they would be better positioned to interact with students. Usually a pair of teachers conducted the instruction with a class of students while colleagues observed and interacted with small groups of students to characterise student thinking according to sub-levels of one or more constructs. Teachers used the TOTs system to record evidence of student thinking. During the classroom lesson, participants could interrupt or ‘pause’ activity as needed to draw attention to an unexpected development in student thinking or to propose an alteration in the plan of instruction. During the debriefing sessions that followed, teachers characterised examples of student thinking with respect to the constructs, often displaying samples of student work or replaying instances of student learning. Constructs became tools for dialogue as teachers developed their implications for current and future instruction. Teachers often concluded with plans for future instruction (‘next steps’), and/or for modifications to instruction to be enacted in the near future with other classes at the same or other grade levels.

Mathematical investigations

A second form of community building involved group inquiries about the mathematics of measure. For example, teachers investigated properties of dynamic measures of space, such as how a length can be viewed as motion along a path, area as generated by a length moved through a second length, volume as generated by an area moving through a length, and an angle as a directed rotation. They also considered how to help make fractions such as $\frac{7}{3}$ more intelligible to students, and how measurement can be employed to interpret arithmetic operations with fractions, especially multiplication and addition. These investigations were most often conducted in response to teacher requests during summer institutes, but were also a component of many of the learning labs.

Auditing evidence and communal looks at student learning

At the end of the school day during monthly meetings, we jointly examined evidence of student learning that was being generated by teachers, with an eye toward establishing a trail of evidence so that others could access the basis of evidence for a particular assignment of a student to a construct. We compared this process to auditing a tax return. We also used TOTs to consider progress in student learning at grade levels and across grade levels, so that we could visualise school-wide patterns of development. These visualisations instigated conversations about the aspects of instruction that needed further attention. In addition, during these conversations teachers recommended changes to conceptual tools and the TOTs.

Revisions of conceptual tools and TOTs

We engaged in iterative refinement of lessons by adding ‘teacher notes’ that clarified the instructional intent of tasks and served as guidelines to productive ways of supporting student learning. As noted previously, these were informed by our work together in learning labs. Similarly, as teachers conducted formative assessments, we relied upon the responses to generate guides that a teacher could use to lead productive classroom conversations based on student responses. These guides were subsequently included in lessons. Visualisations and related capacities of TOTs were expanded as teachers used the tool and conversed about progress in student learning during after-school meetings. For example, we added the heat map (Figure 5) and a history function to TOTs to enable teachers to visualise change during the year at multiple grain sizes (class, grade, school). Constructs were revised to include narratives of development, so that teachers could more readily interpret the progress mapped in the tables.

Evidence of teacher and student learning

There are multiple sources of evidence for the robustness of this innovation at different levels of organisation, ranging from district/school to individual teacher and student.

School level

At an organisational level, the innovation is now part of the school's yearly improvement plan and is endorsed by the district as a resource for K–6 mathematics instruction. The building principal has changed, but administrative support for this innovation remains solid. Teacher participation has remained steady with a few additional participants joining during the course of the project. Teacher corner contributions continue to grow, and teachers have insisted on maintaining the learning lab and mathematical investigations components of the community-building enterprise. Statewide summative assessments now suggest that the school is achieving 'value added' in mathematics, especially for those portions of the assessment indicating measurement and rational number.

Teacher level

To gauge growth in a shared professional vision about teaching and learning measure, we conducted flexible interviews on a yearly schedule to inquire about what teachers notice as they observe videotaped lessons about measure, and about their interpretations of the different forms of activity in which they are engaged. We also examine records of learning labs, mathematical investigations, and formative assessment conversations for evidence of growth of professional vision. As an example, we briefly describe change after one year of participation in the professional learning community in what teachers noticed about instruction in measurement.

At the outset of our collaboration with teachers in Sleeve Elementary, teachers viewed three episodes of classroom teaching in measurement. The teaching episodes were drawn from Grade 1, Grade 3, and Grade 5 and were conducted by teachers from a previous research project that investigated longitudinal change in student thinking about measurement. We asked teachers to tell us what they noticed (Sherin et al., 2011) about concepts of measure and about instructional practices with the aim of exploring the growth of professional vision. We solicited teacher noticings again at the end of the first year of our collaboration.

On both occasions we transcribed video and identified segments during which teachers noticed a core concept of measure and/or an instructional practice aimed at fostering student learning. Three overarching classes of codes were employed to characterise what teachers noticed. The first, *Measurement Concepts*, characterised which concepts of measure that teachers tended to notice, such as the need to define an attribute in one episode and the use of dissection to find area measure in a second episode. The second class of code, *Domain-Independent Practices*, described teacher noticings of instructional practices that supported student learning generally by fostering a positive classroom climate. For instance, a participant might mention that the instructor in the video episode encouraged students to share solution strategies or to ask questions. However, these practices were not explicitly related to learning any concept of measure. In contrast, the third class, *Concept-Specific Teaching Practices*, were forms of instructor practice described as helping students learn specifically about one or more of the core concepts in measure. For example, a participant might notice that the instructor employed a metaphor of motion (e.g. travelling a distance, sweeping a length through another length) to help students differentiate between area and perimeter.

We focused on significant transitions between the first and second interviews, which were given one year apart, in what teachers noticed about core concepts and instructional practices. We counted every instance of teacher noticing about instructional practices across all three of the

video episodes. At the outset of the project (first interview), teachers most often noticed domain-general practices, which accounted for 54 per cent of noticings about instructional practices. These included instructors' questions ('they are using questioning, and the questions I see were ... those higher-level questioning techniques'), instructors' support for student agency ('encourage other students to build upon the thinking of another child'), and instructors' use of materials to support student learning (e.g. 'They are using a lot of visuals'). In contrast, at the second interview, domain-independent noticings decreased to 13 per cent of the total noticings of instructional practice. But noticings of concept-specific instructional practices increased by 61 per cent. And noticings of core measurement concepts increased by 28 per cent, suggesting that teachers were becoming more attuned to coordinating instructional support with identified domain-specific conceptual goals. Table 1 illustrates change in teachers' interpretive framework across all three of the video episodes that they viewed.

Table 1 Transitions in teachers' interpretive frameworks

Concept/practice noticed		At the onset	One year later
Episode 1 Directly comparing heights and girths of pumpkins			
Grade 1 students compare the lengths of paper strips generated by different small groups to represent the height of the same pumpkin.			
Concept:	Define attribute	91%	100%
	Direct comparison	18%	82%
	Origin of measure	9%	64%
Practice:	Highlight variability	45%	91%
	Problematiser comparison	9%	100%
Episode 2 Finding area and perimeter of an irregular polygon			
Grade 3 students considered how to find the area and perimeter of a C-shaped polygon figure.			
Concept:	Unit	91%	73%
	Properties of a rectangle	64%	73%
	Dynamic generation of length and/or area	36%	55%
	Differentiation between area and perimeter	36%	55%
	Dissection of area	18%	82%
Practice:	Highlight defining properties of a rectangle	45%	55%
	Appeal to dynamic motion	27%	82%
	Annotate figure	9%	82%
	Gestures to support learning	36%	82%
Episode 3 Interpreting the meaning of a formula for volume measure			
Grade 5 students interpret the meaning of a familiar formula for the measure of the volume of a prism: length x width x height.			
Practice:	Appeal to dynamic motion	55%	91%
	Tangible model supports visualisation of unit, composite unit (layers)	55%	100%
	Elicit student drawings	18%	91%
	Highlight unit	27%	55%
	Problematiser comparison	36%	100%

Student level

The evidence for student learning includes students' responses to summative, construct-based assessments at the beginning and end of every school year. In addition, the predominant form of evidence consists of evidence generated in classrooms of construct-centered growth in conceptions of measure in length, area, volume and angle. This growth is evident in timeline views of heat maps within classrooms and grade level during the year. For example, a timeline view of Mr. M's first grade class during the second year (2019–2020), displayed in Figure 7, can be interpreted as initial understandings by students of the role of measurement in comparing attributes and properties of units, such as tiling (November). The next snapshot indicates an important conceptual transition to understanding unit iteration (3A) and symbolisation of units (e.g. 0, 1) on a scale, (3B, C), by mid-year, and then further progress toward part-unit iteration (4A) and location of part-units on a scale (4B) by early spring (March in the Northern Hemisphere). Further evidence of learning was interrupted by school closure due to the COVID-19 pandemic.

Figure 7 History of learning about length measure in a Grade 1 class

Discussion

Fostering practices of assessment so that they serve as routine guides to teaching and learning is a goal of most programs of ambitious instruction in mathematics. Knowledge of student thinking and of typical horizons of change are repeatedly cited as critical components of teacher knowledge that undergird adaptive instruction (e.g. Copur-Gencturk et al., 2019; Gibbons & Cobb, 2017). Yet even though teachers' ongoing assessments of student learning are vital to instruction, they are not routinely incorporated into systems of assessment that are used for accountability purposes. To do so, we have identified a set of resources that we believe are vital for bringing teacher voice to larger-scale, summative assessment. One resource is organisational – the need to institutionally support continued teacher learning and collaboration. In this project, we have adapted the math lab approach to continuous improvement of teaching and learning so that assessment practices become strongly coupled to student (and teacher) learning. Instruction is informed by continuous formative assessment, with an expanded sense of what constitutes an 'item' in the traditional sense of assessment. Of course,

this kind of continuous assessment would not be possible without tools like TOTs, which afford capture of student thinking and visualisation of progress at multiple levels of inquiry.

A second resource consists of a common language of learning that can be employed to interpret student responses in a variety of settings. In this project, these are manifested as constructs, which are representations cast at an intermediate level of description. The level of description is chosen to be noticeable as professional vision (Goodwin, 2018) develops, and to be actionable, in the sense that the construct description of student thinking is specific enough to warrant instructional support. Instructional support is assisted by curricular tasks and tools, especially as these are deployed during learning labs. The ensemble of curricular co-design, routine practice of formative assessment embedded in ongoing classroom activity, and a community of practice support children's and teachers' learning (as well as those of us from the university).

The fact that they are designed with common constructs in mind does not necessarily imply that student performances on summative tests and in classroom tasks will be identical. We do not conceive of students as having or not having a particular property that is being measured, but instead think of students as manifesting particular understandings in particular settings. That is, measurement of qualities of thinking is entangled with the circumstances of its generation. What we anticipate is that with constructs, we can interpret student responses consistently across settings and tasks, taking into account variation in circumstances of performance. We are still in the midst of this innovation, so more definitive relations between summative and classroom assessment are still being investigated.

The research reported was conducted in collaboration with Leona Schauble, Corey Brady, and Panchompoo Wisittanawat, and with our K–5 teacher partners at Sleeve Elementary and Mallard Elementary. Funding was provided by the National Science Foundation, DRK-12 program, # 1621088.

References

- Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as a learning profession: Handbook of policy and practice*. (pp. 3–32). Jossey-Bass.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability*, 21, 5–31. <https://doi.org/10.1007/s11092-008-9068-5>
- Copur-Gencturk, Y., Plowman, D., & Bai, H. (2019). Mathematics teachers' learning: Identifying key learning opportunities linked to teachers' knowledge growth. *American Educational Research Association*, 56(5), 1590–1628. <https://doi.org/10.3102/0002831218820033>
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. <https://doi.org/10.3102/0013189X08331140>
- Gibbons, L., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of Teacher Education*, 68(4), 411–425. <https://doi.org/10.1177/0022487117702579>
- Goodwin, C. (2018). *Co-operative action*. Cambridge University Press.

- Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, 52(2), 208–242. <https://doi.org/10.3102/0002831215573773>
- Kazemi, E., Gibbons, L. K., Lewis, R., Fox, A., Hintz, A. B., Kelley-Petersen, M., & Balf, R. (2018). Math labs: Teachers, teacher educators, and school leaders learning together with and from their own students. *NCSM Journal of Mathematics Education Leadership*, 19(1), 23–36.
- Kim, M. J., & Lehrer, R. (2015). Using learning progressions to design instructional trajectories. In C. Suurtamm & A. R. McDuffie (Eds.), *Assessment to enhance teaching and learning* (pp. 27–38). The National Council of Teachers of Mathematics.
- Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is not a stage. *Science Education*, 99(3), 432–437. <https://doi.org/10.1002/sce.21168>
- William, D. (2015). Assessment: A powerful focus for the improvement of mathematics instruction. In C. Suurtamm & A. R. McDuffie (Eds.), *Assessment to enhance teaching and learning* (pp. 247–253). The National Council of Teachers of Mathematics.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.