Using Learning Trajectories to Personalise and Improve Math(s) Instruction at Scale

Jere Confrey
North Carolina State University
<jconfre@ncsu.edu>

Meetal Shah
North Carolina State University
<mshah2@ncsu.edu>

Emily Toutkoushian[1]
North Carolina State University
< ektoutko@ncsu.edu>

This paper discusses critical elements of a dynamic model for taking learning trajectories (LTs) to scale via digital learning systems (DLSs) to deliver diagnostic assessments that produce data on students' progress along LTs for students and teachers during the course of instruction. The paper outlines fundamental qualities of LTs based on design studies, and reviews literature on their measurement and use. It reports on 3 years of design, implementation, and validation of Math-Mapper 6-8, an LT-based DLS as a model to scale LTs. It leverages Lehrer's (2013) notion of a "trading zone" to describe how a learning-trajectory based DLS requires a system of continuous trading among learning scientists, practitioners, and psychometricians. The concept is illustrated using single case of an evolving LT whose focus is on Qualitative Graphing.

Internationally, policymakers have expressed increasing interest in leveraging learning trajectories (LTs) to improve instructional practices and increase student learning at scale in the United States, Australia, and elsewhere. In the United States, the Common Core State Standards (CCSS-M) (CCSI, 2010) were built with consideration of underlying research on LTs (Confrey, Maloney, & Nguyen, 2014; Daro, Mosher, & Corcoran, 2011; McCallum, 2015). In Australia, the Gonski report, "Through Growth to Achievement: Report of the Review to Achieve Educational Excellence in Australian Schools" (Gonski et al., 2018), called for a focus on student learning as measured by growth and states:

> To achieve this shift to growth, the Review Panel believes it is essential to move from a year-based curriculum to a curriculum expressed as learning progressions independent of year or age. (p. *x*).

More recently, attention to learning trajectories/progressions (LT/LPs) is arising in the context of international testing. Adams, Jackson, and Turner (2018) argue that LPs offer 'innovative middle ground' between the rigidity of a single international tests (e.g., PISA, TIMMS) and flexible support of local interpretations of achievement of shared and sustainable goals.

An OECD report reviewing research on LT/LPs in mathematics argued that

> …the discrepancy between what one's students know and what is slated to be taught places teachers in a dilemma that has severe implications for student learning: either teach the official grade-level expectations of the curriculum or address students' learning needs. (Confrey, 2019, p.3)

Confrey (2019) proposed a model by which LTs could play a pivotal role in addressing that dilemma, at scale, by embedding them in a dynamic, digital, psychometrically valid learning system that provides timely and relevant diagnostic information to teachers, students, and parents. Doing so requires three fundamental commitments: a) founding the system on LTs with a rich empirical base on how students reason and draw on diverse cultural resources while solving increasingly challenging tasks; b) measuring progress along LTs using varied tasks associated with levels that can produce systematic, comprehensive, and valid data on learning progress over time, and c) implementing that system in classroom

[1] We wish to acknowledge Dr. Alan Maloney for his editing contribution.

2019. In G. Hine, S. Blackley, & A. Cooke (Eds.). Mathematics Education Research: Impacting Practice *(Proceedings of the 42nd annual conference of the Mathematics Education Research Group of Australasia)* pp. 15-35. Perth: MERGA.

practice, taking into consideration varied and evolving components of curriculum, instruction, assessment, and professional development (PD) to achieve equitable outcomes. She argued this endeavour must be situated in networked learning communities (Bryk, Gomez, & Grunow, 2011) that regularly use data to improve instruction.

This paper reviews fundamental features of LTs as modelled by learning scientists engaged in design studies. It proposes that a diagnostic learning system (DLS) incorporating LTs and measurement of those LTs can provide a means of going to scale if used in the context of diagnostic classroom assessment. We argue that successful scaling requires a means to anticipate varied and evolving LTs across contexts and over time. By classroom assessment (Pellegrino, DiBello & Goldman, 2016), we refer to assessments given during the course of instruction, designed to provide feedback for personalised instruction (in contrast to end-of-unit testing, benchmark, or summative assessments). To illustrate a DLS designed for such a purpose, we introduce Math-Mapper 6-8 (MM) and describe how the processes of its design, implementation, and validation have drawn on extensive collaborations among learning scientists, psychometricians, and practitioners. We present a description of that process for one LT on Qualitative Graphing to illustrate critical elements of continuous collaboration required in successfully taking LTs to scale.

Interpretations, Measurement, and Use of LTs

As with all scientific developments, the origin of the idea of LTs shapes their evolution. The concept of (hypothetical) LT derives from Simon's (1995) interest in describing how an idea evolves within interactive classroom practice as teachers plan to introduce a concept and build on student thinking: the teacher envisions an LT, a hypothesis of the evolutionary states of knowledge the class will traverse as students' reasoning becomes increasingly sophisticated. We use the term evolutionary in the sense of genetic epistemology, a theory that postulated a need for a gradual building up of an idea through assimilation, accommodation, and reflective abstraction in response to a sequence of tasks of increasing challenge (Piaget, 1970). Simon recognised that in this process the "path" had to be hypothetical in that its actual evolution would result from negotiations among students and the teacher drawing on peer-to-peer student interactions and leveraging emergent ideas (Gravemeijer, 1999). Simon's characterisation of the HLT was bound to a goal of enacting what Confrey et al. (2017) have come to define as "learner-centred instruction".

In summarizing the research on LTs, Confrey (2019) reported on variations in the use of the terms progressions and trajectories, attributing the most frequent use of progressions to work by science educators (Corcoran, Mosher, & Rogat, 2009; NRC, 2007; NGSS Lead States, 2013). Though some have argued for a firm distinction between progressions and trajectories (Battista, 2011; Clements & Sarama, 2004), Confrey suggested that the two terms and their meanings have been intermingled (e.g., in the U.S. context where CCSS-M writers' chose to use the term 'progression' in standards). Therefore, we use learning trajectory/learning progression, with the abbreviation LT/LP.

It is important to first clarify what LT/LPs are and are not. LT/LPs are based on empirical research on student thinking and have their roots in a particular kind of design study in which students are provided opportunities to reason with rich tasks that connect to their experiences and that challenge them to actively build, share, and discuss ideas about the target construct. Thus, LT/LPs are always domain-specific and fundamentally linked to instruction and describe obstacles and landmarks encountered by learners as they move from naive to sophisticated reasoning about target constructs over time. LT/LPs are comprised of epistemological objects including naive and partial conceptions, limited and/or multiple representations, strategies, explanations, cases, emergent properties (formal and informal) and generalisations. Logico-mathematical deconstruction can inform

development of LT/LPs, but it is only a starting place. Finally, LT/LPs represent expected probabilities, not stage theories, and can evolve and vary to a degree as they are influenced by student experience, cultural tools and practices, and language and instruction. In our work, even though a trajectory is often displayed as a ladder of progress levels, we liken it to a climbing wall, replete with handholds and obstacles, which will be entered and traversed by different learners in different ways (Confrey & Toutkoushian, in press). Nonetheless, the challenges and opportunities students are likely to encounter are often predictable.

Six commitments are shared among most LT researchers: 1) LT/LPs are conjectures that involve modelling levels of learning; 2) their levels progress with respect to a specific domain and are situated in constructivist and socio-cultural perspectives, 3) the transitions between levels of an LT/LP require acts of reflective abstraction to achieve increasing sophistication, 4) instruction plays a significant role in student knowledge growth, 5) multiple ways to traverse the levels are expected, and 6) the emergence and visibility of the LT/LPs will depend on learner-centred instruction. (Confrey, 2019)

The pragmatic value of an LT/LP is that it can help teachers hear and recognise student reasoning and build on it during instruction. Research on the use of LT/LPs in classrooms have shown both promise and challenge. Small scale studies using significant examples of student work have reported that teachers learn to better leverage student thinking (Suh & Seshaiyer 2015; Wilson, Mojica, & Confrey, 2013; Wilson, Sztajn, Edgington, & Confrey, 2014). Other studies (e.g. Supovitz, Ebby, & Sirinides, 2013), employing a coding framework for teacher explanations of student work (ranging from general to procedural, conceptual, and then LT-based reasoning), found a prevalence of procedural reasoning and only minor evidence of LT/LP based explanations. At a larger scale, work focused on specific LT/LPs has shown evidence of positive effects on teacher and student learning (Clements, Sarama, Wolfe, & Spitler, 2013; Lehrer, Jones, Pffaf, & Shinohara, 2017; Siemon et al., 2018).

To consistently achieve benefits of LTs, teachers need methods to systematically and regularly gather data on student progress, and use it formatively, to improve instruction and increase students' self-regulation (Heritage, 2008). Teachers can be provided with formative assessment prompts and learn to analyse the levels of those artefacts of student work (Petit, Laird, Marsden, & Ebby, 2015; Simon, 1995). While such an approach is highly authentic, it can be time-consuming to score and tends to only shed light on some aspects of the LT/LP. Other researchers opt to create measures of student progress along LT/LPs and are applying a variety of empirical techniques (e.g., item response theory, Bayesian networks, latent trait analysis, and cognitive diagnostic models) to model student progress along LT/LPs (Pham et al., 2017; van Rijn, Graf, & Deane, 2014; Wilson, 2012). Some are concerned that applying psychometric models (Battista, 2011; Stacey & Steinle, 2006) will diminish the intended variation among the levels reducing it to a generic difficulty parameter. However, a substantial advantage of using digitally scored measures and applying psychometric models to the data is that the data can be collected systematically from all students, scored, and returned to students and teachers immediately. If assessments are equated and vertically aligned, they can also be used to efficiently chart progress along the LTs (Wilson, 2013).

Administering measures of LT/LPs in the context of a DLS makes it feasible to iteratively design, revise, and validate processes at scale. Such steps are critical to register, and act on our commitments to foster the expected evolution of LTs as well as their measures based on empirical data. However, a principled process of revision of LT/LP constructs and features of the DLS need to be coordinated with the use of the DLS in practice. Such an effort requires a step beyond the already complex process of using data to drive instruction. To envision how to do this, we drew on Lehrer's (2013) use of the term 'trading zone', which

conceptualises LT/LPs as a joint production of learning scientists, psychometricians, and practitioners, each sharing and contesting their disparate expertise and insights.[2] Lehrer's ground-breaking work on data modelling was conducted at a significant but smaller scale than we hope for the use our DLS, which we describe in detail in the next section. Lehrer's initiatives depended on six LT/LPs, intensive interactions with participating teachers, and a shared, albeit novel and evolving, curriculum. In contrast, we have aimed at middle school mathematics (grades 6-8) as a whole (a subject explicitly regulated with standards and high stakes testing), by providing 62 LT/LPs and their measures, limited periodic PD, and the flexibility of use across varied curricular programs. So, a critical question for our project is whether and how the concept of a trading zone can be applied in the more extensive context of our DLS.

Partnering to Use Math-Mapper 6-8

In order to take the use of LT/LPs to scale for classroom instruction and assessment, Confrey (2015) and her research team built a web-based application founded on LTs[3], Math Mapper 6-8 (MM) (sudds.co) for middle grades mathematics. It contains a hierarchically organised learning map with nine big ideas, 25 relational learning clusters (RLCs), and 62 research-based LTs, each associated with CCSS-M. Its diagnostic assessment system measures students' progress along LTs and provides real-time data to students and teachers through user-friendly displays. Teachers embed the diagnostic assessments in instruction as classroom assessments based on their curricular sequence (Wilson, 2018). Assessments are at the level of the RLC and comprise 8-12 items of varying types. Items are aligned to LT-levels and designed to stimulate rich classroom discourse. Data are returned in displays that support teachers and students in reviewing individual items, analysing student response patterns, and alerting them to common misconceptions.

MM has been developed in partnership with three schools[4] in two districts over three years. District 1 is a stronger-performing, wealthier district with more experience with 1:1 digital resources. District 2 serves a more diverse (ethnically, socioeconomically) and relatively transient (military families) population. The research team has conducted annual PD workshops, observed classes and numerous examples of data return and discussion, and studied PLC meetings in both districts. Across those schools, 36,684 assessments have been used in the first full round of validation. This work in total has provided the evidentiary basis for the discussion of the case described here.

Trading in the Trading Zone (Phase 1): Building the LT and its Measure

In the OECD report, Confrey (2019) argued that the path forward to taking LT/LPs to scale required the development of dynamic digitally-driven DLSs informed by a variety of factors, and functioning as a continuous improvement system (Bryk et al., 2011). In this paper, we have described MM as representing a first generation of such a system. We envision that the LTs undergo continuous evolution, more vigorously early in the process and more gradually later on, and with possible episodes of major revisions in light of new research or changes in instructional scope, topic, or materials/tools. We also envision that different LTs could be proposed, then requiring comparison and contrast. This evolutionary

[2]Although Lehrer argued "...learning progressions do not accord well with metaphors of ladders or pathways of development. Instead, they are more like what P. Galison (1997) calls a trading zone" Lehrer, 2013, p. 173), we suggest both (climbing wall and trading zone) are useful metaphors for process and outcome.

[3] As a convention, when referring to Math-Mapper specifically, we use the term learning trajectory (LT).

[4] A fourth school has joined this year, but it is not included in the reported validation data.

process should be negotiated among learning scientists, practitioners, and psychometricians who mediate the influences (both affordances and constraints).[5] Drawing on Lehrer (2013), we view the exchanges among these participants and ways in which groups interact directly and indirectly, a form of a 'trading zone'. In the remainder of the paper, we present, as a case of working in a trading zone, the development of a single construct, 'Qualitative Graphing', within the cluster 'Exploring Relations and Functions', in the algebraic big idea 'Represent and use relations and functions of two variables'. This cluster contains a second construct called 'Relations and Functions'. The case seeks to illustrate how varied types of expertise interact to shape an ever-evolving LT, its measurement, and its use in classrooms. We briefly describe the literature base from which the LT was built and discuss how it is situated in relation to the American CCSS-M.

Research base on learning. Qualitative graphing has a long history in mathematics education to serve as a means to build students' intuition about functions as represented in linking situations to graphical sketches. Tracing back to Vinner (1983), researchers recognised that students' concept image of function was at odds with their concept definition and exerted significant influence on their responses to questions. It was clear that students' understanding of functions as symbolic equations was limited and poorly coordinated across a key set of representations (graphs, tables, and equations). Using everyday common situations and technological tools such as Logo turtles, motion detectors, or simulations, researchers documented that students, even young students, could conceptualize position-time graphs and reason about rate informally (Papert, 1980). Rich simulations were sometimes used to improve such reasoning (e.g., SIMCALC (Kaput & Roschelle, 2013)). Such simulations demonstrated that students could even reason in the context of bank accounts and rain gauges, with descriptions of discontinuous rates of change (transactions, daily rain accumulations), to produce accumulations (Wilhelm & Confrey, 2003) presaging the relationship between a derivative and its integral. Qualitative graphing sessions allowed students to wrestle with the idea of changing rates without having to simultaneously tackle the paradoxical character of instantaneous rate. Researchers further found that by pairing position-time graphs with velocity-time graphs, certain ideas of rate of change and accumulation could presage an introduction to elementary concepts of calculus (de Beer, Gravemeijer, & van Eijck, 2015; Monk, 1992; Roschelle, Kaput, & Stroup, 2000). What was essential among these studies was to bolster a multi-representational understanding of functions and relate it to students' situational knowledge.

Relationship to CCSS-M. In the United States, efforts are being made to link Standards and LT/LPs, but the relationship is complicated. Standards serve the purpose of setting grade level instructional targets and coordinating instruction across grades; while one might propose that the LT/LPs be used as standards, it is actually not feasible. Standards are generally designed to be parsimonious; their grain sizes can vary significantly (Confrey, Maloney, & Corley, 2014), and they are placed at a single grade to indicate when students can be held accountable for knowing the topic. As a result, a standard seldom addresses levels of reasoning prior to its specified grade. While one can align the top levels of an LT/LP with a standard, the ramps to understanding, represented at a finer grain size by the LT/LPs, require distinct articulation. In our work with teachers, we emphasise that the LT/LPs are 'aligned' with standards, and encourage them to focus instructionally on the LT/LPs.

For example, the topic of Qualitative Graphing, as a means of intuitively understanding functions in context, should accompany the introduction to functions and extend across

[5] The software designers and engineers represent an additional influence beyond the scope of the paper.

lower grades.[6] Instead (in the United States), only one 8th grade standard (8.F.B.5) directly references qualitative graphing:

> Describe qualitatively the functional relationship between two quantities by analysing a graph (e.g., where the function is increasing or decreasing, linear or nonlinear). Sketch a graph that exhibits the qualitative features of a function that has been described verbally.

Placing Qualitative Graphing at the 8th grade provides teachers no indication that the topic should be introduced in the primary grades or, at least, earlier in middle grades. Thus, a dilemma for those of us who build LT/LPs is how to accurately reflect the evidence on the developmental timing of learning while respecting the standards' authority to specify when topics are to be taught (and tested) across schools. To address the dilemma in MM, the highest levels of a construct are officially mapped to a target standard. Grades are tentatively associated with each lower level based on research, and if and when a relevant lower grade standard can be found, it too is aligned. The coordination of the research on learning with the country or state's standards represents an important element of the 'trading zone' in that it has a significant impact on how feasible it is for practitioners to use the tool, especially since the standards are used to build high stakes assessments.

Versions of the Qualitative Graphing LT. Table 1 presents the original and revised (based on validation discussed in a later section) versions of the LT. Of the seven levels, two are associated with sixth and seventh grade, and align loosely with standard 6.EE.C which discusses relating graphs, tables, and equations and leaves room for qualitative graphing. All other levels are assigned as 8th grade, which reassures teachers who want to delay the formal introduction of 'slope' to 8th grade. This is done despite recognizing that younger learners can reason informally about steepness.

The LT is sequenced so that students learn to relate changes in position-time graphs with associated movements (towards, away, and at rest) and apply that reasoning to increasingly complex graphs. The LT starts with interpreting a single linear equation in terms of movement along a line from a point (L1) and then shifts to piecewise linear functions (L2 and L3). At L3, the students distinguish among descriptions of position, time and rate or speed, typically associating speed with steepness. At L4, students link the direction of the slant, "going up or down from left to right" (positive and negative slope), with velocity. At L5, we initially predicted that the LT/LP could include linking position-time and velocity-time graphs on linear functions. The last two levels describe students' ability to analyse or sketch curves, which require students to coordinate the idea of continuously changing rates with position-time graphs, first for simple (L6) and, later, more complex, curves (L7). Two misconceptions are associated with this LT. The first is the misconception that a student "Believes the position vs. time graph is the same shape as the path an object travels". The second is that a student "Interprets position vs. time graphs as velocity vs. time graphs".

The LT must foremost be consistent with the levels of sophistication implied by the empirical data on learning. Basic alignment with standards is engineered, more or less. Curricular resources affect the learning sequence as well; for instance, if students have access to the automatic feedback produced by motion sensors, then they are likely to encounter and analyse curves and discuss speed earlier. Thus, the LT represents a hypothetical model designed to generalize across these interacting factors and resources.

Table 1.
A Learning Trajectory for Qualitative Graphing

Level	Original Level Description and Order	Revised Level Description and Order	Gr.

[6]Confrey has observed second graders making sense of position-time graphs using a motion sensor.

1	Interprets changes in position and time qualitatively as moving toward, moving away, or stopping for linear graphs	Qualitatively interprets or graphically represents changes in position and time as moving toward, moving away or stopping	6
2	Interprets changes in direction, but not rate, for piecewise linear graphs qualitatively	Qualitatively interprets or graphically represents changes in direction but not rate for piecewise linear position vs. time graphs.	7
3	Interprets changes in speed (slope) and direction for piecewise linear position vs. time graphs qualitatively	Qualitatively interprets or graphically represents in speed(slope) and direction for piecewise linear position vs. time graphs	8
4	Differentiates speed (magnitude of steepness) and velocity (magnitude and direction) to distinguish negative and positive slopes in context	Differentiates speed (magnitude of steepness) and velocity (magnitude and direction) to distinguish negative and positive slopes in context	8
5	Distinguishes and relates graphs of position vs. time and velocity vs. time	Qualitatively interprets or graphically represents changes in direction and rate for simple increasing/decreasing position vs. time curves	8
6	Interprets continuous changes in rate for basic curve shapes in multiple contexts	Qualitatively interprets or graphically represents changes in direction and rate for complex position vs. time curves	8
7	Interprets continuous changes in rate for complex curve shapes in multiple contexts	Distinguishes and relates graphs of position vs. time to velocity vs. time	8

After the LT levels are described, items are written according to an elaboration document that specifies associated misconceptions, cases, and numeric value ranges for each level. The following item types are used: multiple choice, select multiple, numeric response, 1-letter matching, and true/false are used in MM. To illustrate how the reasoning demand of items increases in sophistication, Figures 1 a, b, c, and d, illustrate levels 1, 3, 6, and 7 of the original LT (1, 3, 5, and 6 of the revised LT). Item analyses are included, based on data from all participating schools.

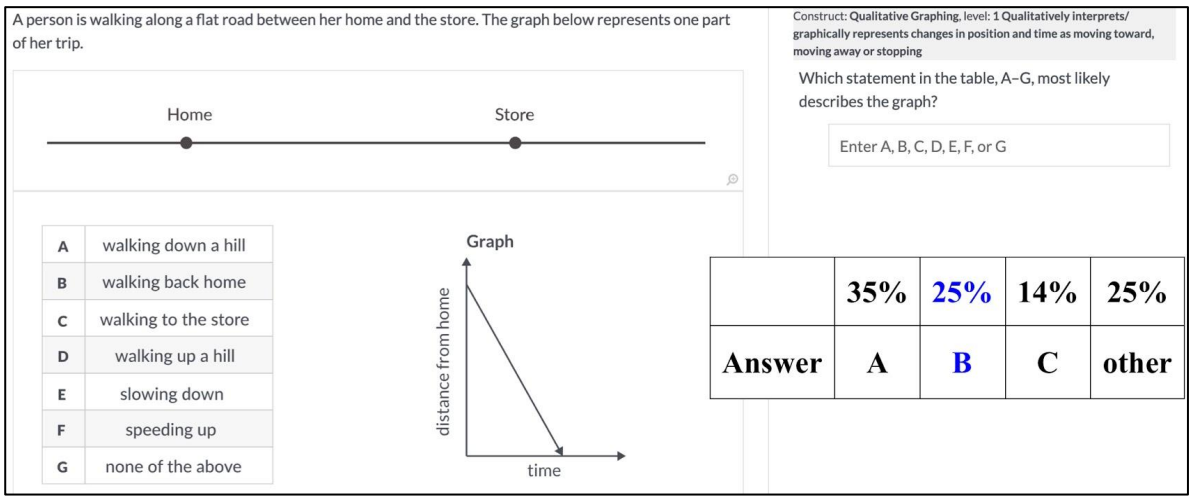


Figure 1a. Item at level 1 of the Qualitative Graphing LT (*n*=361 responses).

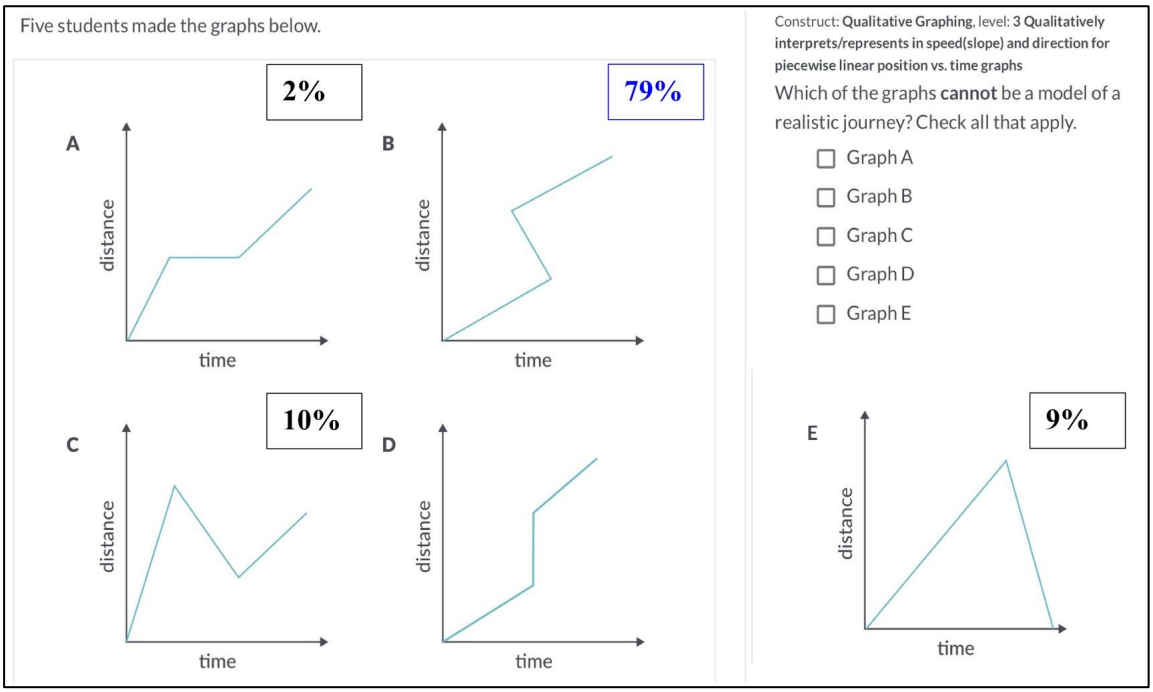


Figure 1b. Item at level 3 of the Qualitative Graphing LT (*n*=346 responses).

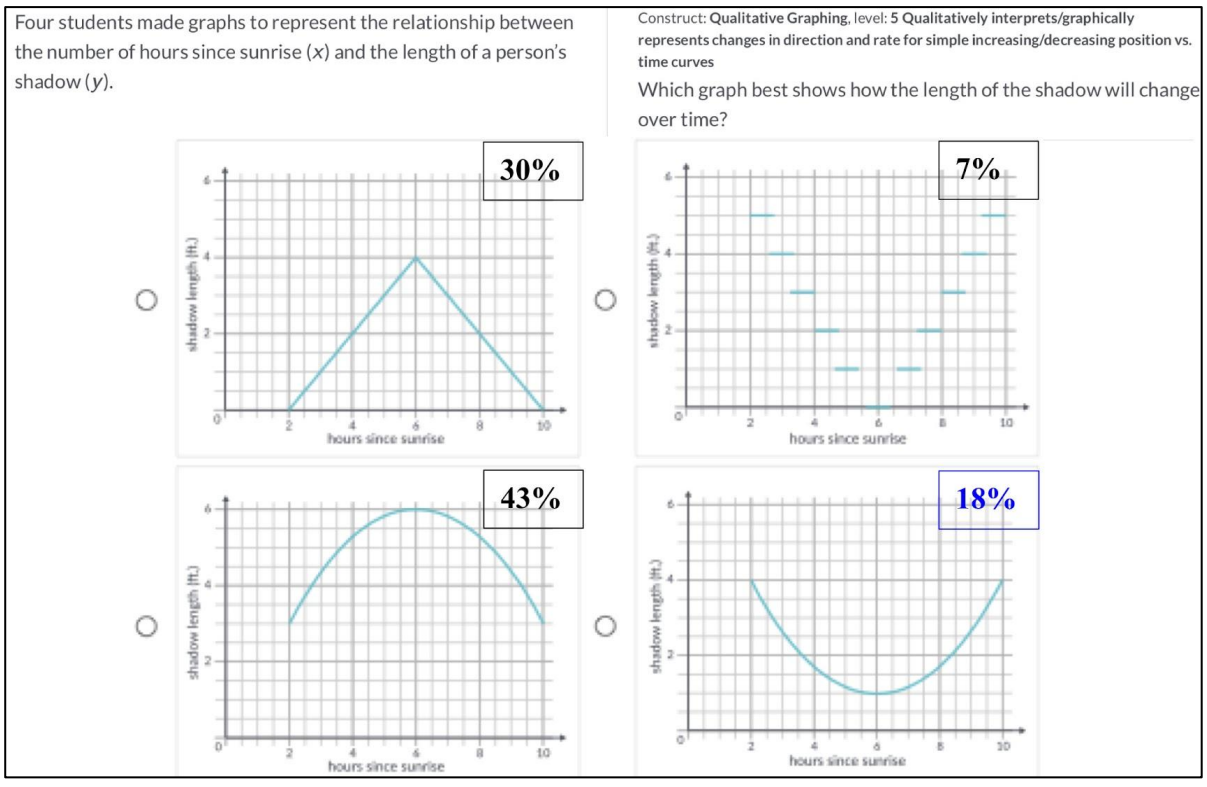


Figure 1c. Item at new level 5 of the Qualitative Graphing LT (*n*=122 responses).

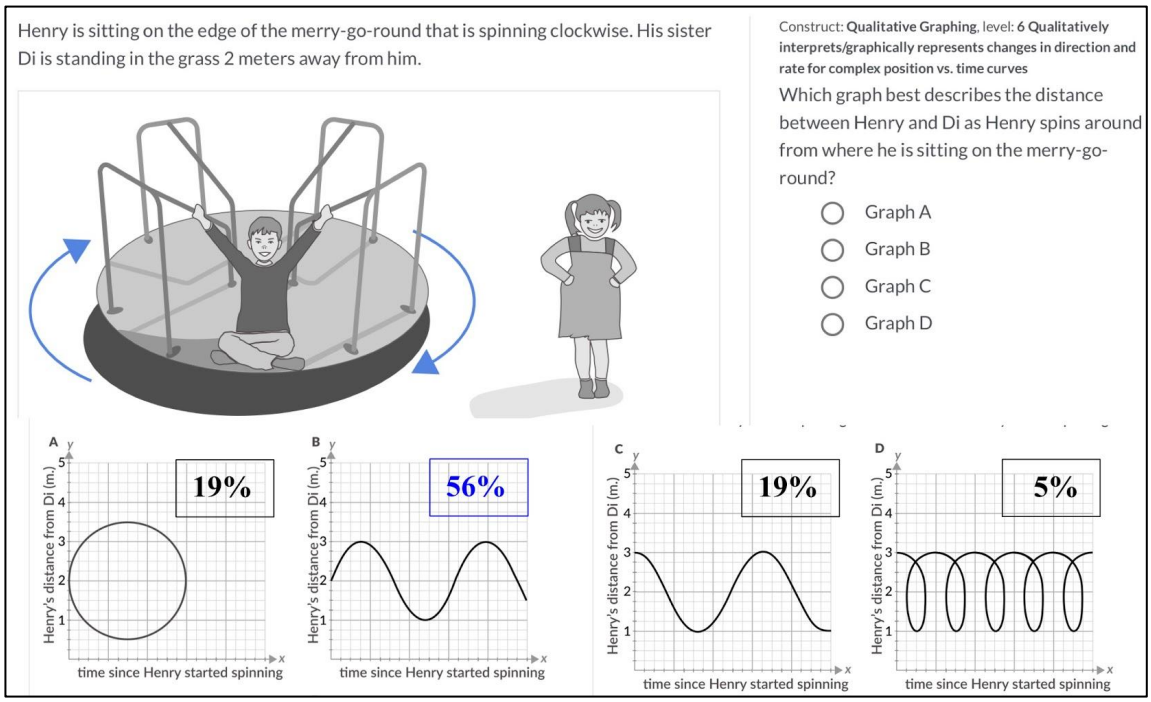


Figure 1d. Item at new level 6 of the Qualitative Graphing LT (n=104 responses).

Trading in the Trading Zone (Phase 2): Implementation and Its Study

Schedule for using MM. Once the LT and related items are constructed, assessments are built for use in classes. One model for using MM in a classroom (Figure 2) is based on the idea that teachers will undertake whole-class instruction on a topic, and complete it with 25-30% of the allotted time remaining for administering a diagnostic assessment, providing personalised instruction, and retesting based on those results. Each assessment samples from levels in all relevant constructs of an RLC. To ensure the assessment covers all levels in an RLC, teachers are provided with multiple, psychometrically equated test forms.

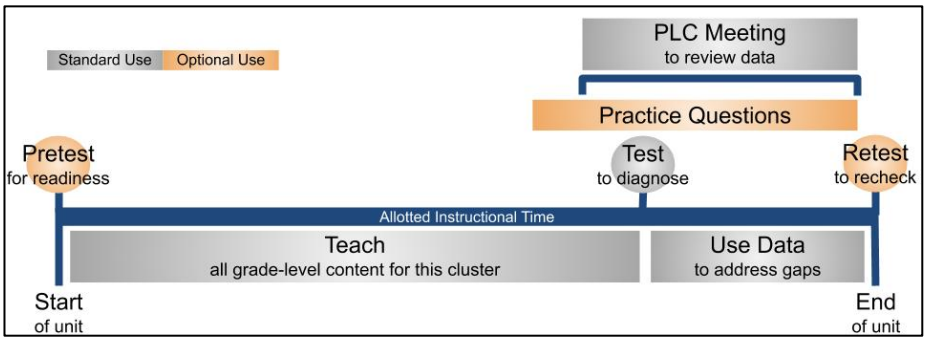


Figure 2. A model for the use of MM.

Data reports are returned immediately to both teachers and students. Students' reports show the percent correct for each construct, and an item matrix that provides more detailed information relating items to constructs and levels. They can access those items and revise (or reveal) the answer to any missed items. If correctly revised, scores improve. Students also have access to the 'Practice mode', in which they can select a construct and level(s) at which they need to practice. Immediate feedback is provided item by item. Teachers receive class-level feedback in the form of a 'heatmap' for each construct in an RLC (Figure 3). This heatmap is from a seventh grade accelerated class which is also discussed in the next section.

Each cell represents a student response to a single item (blue: correct; light blue: partially correct; orange: incorrect). Each row represents a level of the construct (low to high levels from bottom to top). Each column represents a student's items' responses for all levels tested. Students are ordered by their total scores from low (left) to high (right).

Teachers look for a Guttman curve which separates the map into mostly incorrect and mostly correct responses. The two rectangles (Construct B) identify groups of students who need additional help across all levels (a) and sets of levels needing to be revisited with the entire class (b). For example, in Figure 3, the class shows strong mastery construct A. This conclusion is somewhat limited because Levels 4 and 5 were not tested. In contrast, the map for Qualitative Graphing shows only a modestly strong Guttman curve with Level 1 performance atypically weak for an LT. After reviewing the heatmap, teachers can assign an equated re-test within MM and students can initiate a re-test or practice on specific levels.

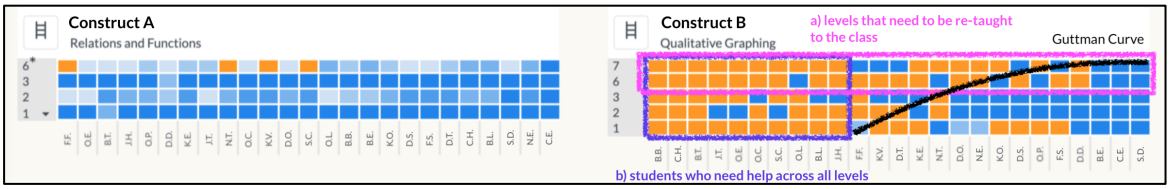


Figure 3. A heat map from a 7th grade Advanced Math class.

From the heatmap, a teacher accesses an item at a level during classes to review the data. In response to our partner teachers, the team added a means to: a) reveal the correct answer, b) see what percent of students show evidence of a given misconception, and c) display an item analysis of student responses. The trading zone metaphor encompasses opportunities for practitioners to indicate resources that would better serve their needs.

Teacher data review. Once students and teachers had access to the trajectories, assessments, and data reports, our focus turned to an important region of the 'trading zone'- the opportunity to study and learn from teachers as they reviewed data with students. Data reviews reveal how students are invited—and respond—to the data-driven approach, and how well that approach fits the context of curricular use and instruction. A study of data return practices (Confrey et al., 2018) documented that high-quality data reviews by teachers exhibit a variety of characteristics. They are learner-centred [LC], promote a growth mindset. [GM], develop students' meta-cognitive awareness [MCA], emphasise salient item features [IF], and discuss patterns in item response [IR], such as misconceptions [MIS]. Weaker reviews occur when teachers fail to leverage the LTs, simply reteach an item the same way, and thus provide little or no opportunity for students to participate.

The transcript below from 7th grade accelerated class includes the codes for quality data review characteristics. This excerpt allows consideration of how this teacher successfully reviews the use of the tool, discusses the data, and communicates her expectations to the class. The teacher first projected the class's heatmap for the cluster (Figure 3) and reviewed how to read it. She asks students to list "two things they think the heat map shows" and the whole class discusses those responses. The first item discussed is shown in Figure 1a (at the time, item 1a included a second graph describing a person traveling to the store).

> T: So, it depends on, um, how you scored in each, what we call, a construct. What we did was we took a test on the cluster of exploring relations and functions. What did you notice looking at this, we call this a heatmap? What did you notice? [MCA]
>
> [Students comment on how well they did as a group on the first construct.]
>
> T: Give me another observation that you made, <Student>
>
> S: The class better understood Relations and Functions than Qualitative Graphing.

T: You got it, you got it, OK. So, what does that tell me? As a teacher? What does that tell me that needs to be done, or what? <Student>

S: Um, you need to work on Qualitative Graphing with us.

T: Right, as a group, we need to maybe reteach that topic, spend more time on it, [GM] Ok? Is there any specific number--and that number is a level--is there any specific number that stands out to you, as one that was like "ooh" not so good? [MCA]

S: Level 1

T: [I] can go right to the question. So, this is what we're gonna look at now, as a class and let's see if we can make some sense out of it. Someone want to read it? Go ahead, <student>

S: A person is walking along a flat road between her home and the store. Graph 1 [figure 2a] and graph 2 [not shown] represent two different parts of her trip.

T: So which statement in the table, A-G most likely describes graph 1? Now before we look at this, what do you think is a common misconception here? [MIS]

S: Because of the slope, you might think it's like going down hill [graph 1] or going up a hill [graph 2].

T: Exactly. Someone's going to say, "I'm going down a hill", "I'm going up a hill" Is that true?

Students: No.

T: Why not? What tells you that? <Student>

S: Because none of the axes say height.

T: They don't say height [LC]. What do they say? [IF]

S: They say, "Distance from home" and "time".

T: "Distance from home" and "time". So, if we look at graph 1: At time 0, where is she? Or he, she? [LC]

S: The farthest away from home.

T: The farthest away from home. So where could she be, according to this problem then? <Student>

S: At the store.

T: At the store. So where is she by the time we hit the *x*-axis?

S: At the store.

T: At the store? She's at the store here, right? [LC]

S: Yes, she's home.

T: She's home. How do you know she's home?

S: Because it's the *x*-axis. [LC]

T: So, it means the distance is what? The distance from home, at that point is…?

S: Zero.

T: So, did she go down hill? [MIS]

S: No.

T: No. It said that she walked on a flat road. Okay?

[The class reviews the second graph and displays the item analysis]

T: So, the good thing is, you can reveal the answers----. Now, what did we answer as a group? That's when I can go here, and it doesn't necessarily give me exact people's names, which is a good thing. But, it tells me how many, and the percentage of the class that actually chose the correct answers [IR]. So, I got, "slowing down", "walking down a hill", was pretty popular, but we talked about that [MIS]. So, when you see a question like that, what is your first thought to do? [GM] Prior to today. When you saw that, because you took this test already. What were you thinking? [LC]

S: That we should probably read the question first. [MCA]

T: Probably read the whole question, and we kind of talked about that yesterday as a group of teachers. We discussed, the fact is, sometimes we read the first sentence, and then we stop. Read the whole question. What else should you do, other than reading these words, what else should you do? <Student>

S: Uh, pay close attention to the words, because, like "flat" I didn't even realize it was there.[IF]

T: You got it. Because you probably might have skipped over it. It's a flat surface, Ok. What else, <Student>?

S: You should look at the labels for the *x*- and *y*-axis.

T: Look at what the independent and dependent variables are. Look at what they mean. .[IF]

S: And you should look at like the graph, the way it's going.

T: Right, is the graph… How do we read a graph?

S: *x*, *y*

T: *x*, *y*, Another way? How do you read a sentence? In which direction?

S: Left to right.

T: Left to right. How do you read a graph? .[IF] Left to right. So, as time increases, what's happening to the distance from home, and then making your interpretation from there. Questions? Alright, so do you think if reviewing this first and then taking the test, the scores would be different? [GM]

Students: Yes.

This transcript illustrates the importance of the data return process. Effective teachers employ the process for feedback on correctness, instructional opportunity to rethink the items, and as a means to increase student metacognitive awareness and agency.

Observation of classroom data reviews provide insights beyond the student response results themselves, sometimes contributing to decisions to stabilise *or* modify LTs. In one instance, we preserved the placement of a level concerning percents above 100--despite atypically weak performance relative to its neighbouring levels--because data review and subsequent review in professional learning communities (PLCs) revealed that the topic had been instructionally overlooked (Confrey et al., 2018). In other instances, data reviews have led to revisions of the LTs or the map structure. Aligning to standards facilitates accurate grade level assessment, but aligning assessments with the curricular grouping of topics during the academic year also represents a necessary practical constraint. MM supports flexible sequencing of clusters during the year, but grouping of topics in clusters for testing is predetermined. Based on teacher feedback and data review observations across multiple sites, we revised the map to place "Positioning and operating with integers" into its own cluster, separating it from the original clusters for ordering and operating with fractions, decimals, and integers: even though all these topics are conceptually related, instruction on integers is often temporally distinct, so map structure and assessment timing needed to reflect the partners' instructional plans. Working in the 'trading zone' with LTs provides opportunities for such practical adjustments and more informed interpretations.

Teachers' individual and collective use of MM. We convened and studied grade-level PLCs in which one teacher from a grade-level team reviews data from her/his heatmap to compare with colleagues' classes' data. The results varied considerably across classes and across teachers, including some consistently stronger or weaker teachers or classes. The collective PLC-based data reviews provided teachers opportunities to share successful approaches. In the short term, they exchanged instructional strategies, and they jointly planned longer-term curricular modifications. The PLCs also provided teachers with opportunities to propose adaptations and extensions to MM. For example, during PLC

sessions early in MM's development, it was teachers who proposed including misconceptions and item analyses.

Capable supervisors often made the difference between effective and ineffective PLCs and data reviews. These administrators prioritised the use of data for improvement and used data return to set concrete and ambitious targets for student growth. They focused discussions on the detailed information in the reports. Weaker supervisors either did not attend the meetings or communicated little accountability and missed opportunities to strengthen teacher content knowledge and convey high expectations for all students.

Another notable contrast among sites concerned how low scores were managed (often 30-60 percent correct per RLC). It must be noted that for an assessment to be diagnostically useful, the range of scores must be centrally distributed. Some teachers emphasised to their students that the scores were not for grades but rather were starting places to indicate where they should focus their efforts. These tended to be teachers who fostered growth mindsets for students and used the problems as opportunities to raise the bar conceptually. Their students were more likely to learn to step up to the conceptually challenging items. For other teachers, low percentages correct caused tension; they were often surprised or defensive. Some reacted by rejecting the measures as not aligned, too difficult, not consistent with their instructional methods, or discouraging to students. These reactions are leading to consideration of different ways of reporting scores.

The low score ranges raise the question of whether the scores should be so low given that they test empirical established patterns in conceptual understanding associated with the grade level expectations. Research on middle grades instruction documents, and our observations confirm, that excessive procedural instruction is rampant, and is often reinforced by the use of almost exclusively procedural materials (Dysarz, 2018). In these settings, students and teachers become accustomed to administering tests that produce inflated scores.

MM was more enthusiastically implemented by teachers in our wealthier district, which presented a unified vision, clear leadership, and a focus on conceptual learning. Effective use of MM was slower and sporadic in partner schools serving more diverse student populations—these tended to display more weakness in instruction. However, during our three-year partnership with the school serving a more diverse population, we have seen major improvements. For the 6th grade, research relating state-wide, summative end-of-year (EOY) test scores to the use of MM shows a significant relationship between students taking additional MM tests and growth in EOY test scores, as well as between teachers administering more tests and average class-level EOY score improvement.

Our ongoing partnerships have helped us to recognise that more forms of scaffolding are required for success. For teachers to understand and trust the LT/LPs, significantly more PD than anticipated is needed, especially working at scale, across 62 LTs in 25 clusters. Our conjecture is that we need a learning trajectory for *teachers* to learn to use MM, understand the LTs, and implement its implied instructional changes. Efforts are underway to develop a phased introduction to MM for teachers.

These results both illustrate and challenge the 'trading zone' concept when the use of LTs is extended to a larger scale. The implementation of teacher-proposed affordances established the value of partnering with schools in ongoing development. The success in implementation at a wealthy, better resourced school served to validate the feasibility and value of the approach and provide clear targets for how to implement well. Observing innovations in that setting have been widely disseminated to the benefit of practitioners.

Our mixed success in implementation with other partners raised questions of how to characterize partnership with all teachers as a 'trading zone' when working at scale. An effective innovation must both fit into enough to teachers' practices to be sustained, while

disruptive enough to lead to transformations of that practice. To enact change at scale often requires some combination of sheer persistence through initial periods of discomfort, a pilot group to lead the adoption, sustained obligatory PD, and commitment, expectations, and accountability from school leaders. If the concept of a trading zone conceived of democratically equivalent decision-making by all participants, it is unlikely that difficult, and yet potentially transformative innovations will take hold. Yet, helpful feedback can come equally from eager and reluctant participants alike. So, a more complex view of participation in the trading zone by practitioners, and a means to include students' feedback, is required. Consultation with trusted partners who have demonstrated shared, fundamental commitments to learner-centeredness and detailed epistemological aspects of student reasoning has been one productive approach. Strengthening the use of digital means of soliciting ongoing input, commonly done with innovative commercial products, is a possibility. Further conceptualisation of this arena is necessary.

Review of Overall Student Data. Data for entire samples are examined by researchers and administrators in the form of compound bar graphs. We emphasise that these data represent overall student results after initial instruction; if teachers do provide additional instruction using these data, change would only be evident on MM re-tests or end-of-unit summative tests.

The results on Qualitative Graphing for our entire sample are shown in Figure 4. Bars represent a tested level; gaps represent untested levels. Bars are grouped (Groups 1 through 4) by district into grade level and class type (regular or advanced). District 1 assigns students to enriched (regular) and pre-algebra (advanced) classes). Group 1 (District 1 7th grade Pre-Algebra) represents the strongest students, though they may not have tackled the whole LT yet. Group 2 represents all District 2 8th graders who are not taking algebra; Group 3 are District 1's stronger 8th grade students who are not taking algebra, and Group 4 are their regular 8th grade students.

From these data, we note that students did better at lower levels (L1-3) than higher levels (L4-7). The order of difficulty of levels 1, 2 3 seem slightly and consistently reversed from expected. L5 is consistently more difficult than expected. These results will be examined again below in the context of the validation studies. Patterns in the data are consistent across the groups. Group 3 has the strongest results (more blue), followed by Group 2, followed by Group 1, and, finally, Group 4. These results confirm that the content is targeted more to 8th grade at both districts, and that perhaps, for district 1, the Grade 7 Pre-Algebra students either had more opportunity to learn the material or could more easily figure it out intuitively than the Grade 8 students who had not been placed into Algebra.

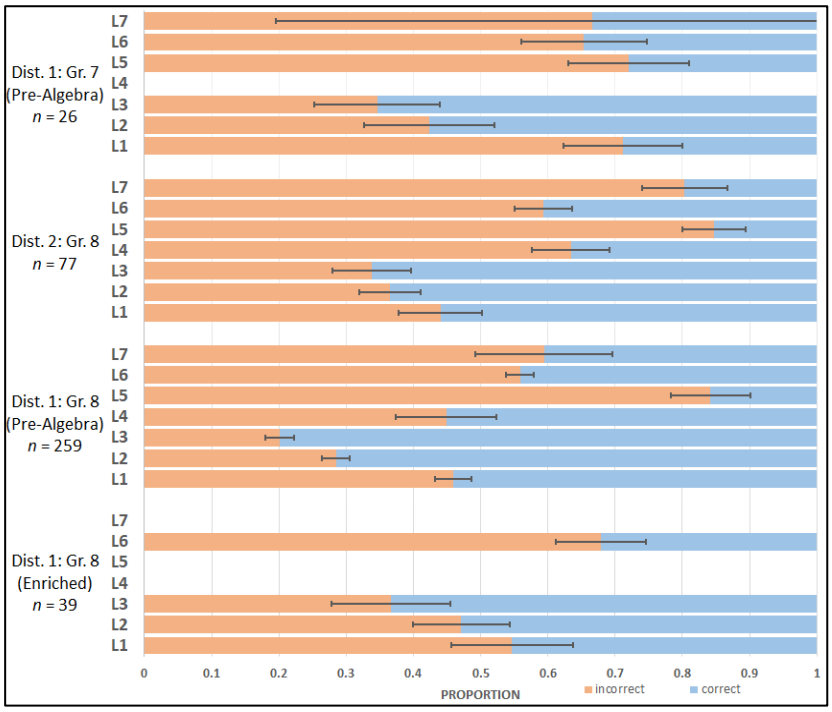


Figure 4. Compound Bar Diagrams for Qualitative Graphing.

While this LT shows some deviation from typical LTs, which show an overall decrease in proportion correct as one moves up the levels, the results confirm that overall the LT gets harder. The data also seem to separate the LT into two halves—Levels 1-3 and Levels 5-7 with Level 4 (not tested in two classes) performing in the middle. This suggests that students are not demonstrating competency in reasoning qualitatively with situations of changing rates and have not fully distinguished velocity-time (or perhaps velocity as distinguished from speed) from position-time. PLC review data seems to indicate the curricular treatment of these topics in the regular course of instruction does not extend to velocity-time graphs.

The role of administrators within a metaphorical trading zone has only been modestly explored. In the context of the PLCs, as mentioned, their role is critical for setting the context for classroom assessment and the use of data. We also know that some teachers fear that poor results will be used to negatively evaluate them; some administrators anticipate this concern and emphasize the importance of using MM data to improve instruction instead of to evaluate teachers. Administrators must set expectations for teachers to act effectively on the results—reteaching, forming subgroups, and assigning targeted resources. It is crucial that administrators set the norm that all students achieve the expected levels of performance, but also support this with sufficient content-specific coaching overseen or delivered by mathematics supervisors, based on data. In order to actually leverage data from MM, districts have to dismantle the silos among their assessment, curriculum, and PD departments.

Trading in the Trading Zone (Phase 3): Validation

Going to scale with a tool that provides assessments for 62 LTs required us to create a suitable continuous validation process. Validation analysis is particularly important early on because of the uneven learning sciences research base. The team developed a validation approach applying item response theory (IRT) to produce item difficulty parameter estimates (Confrey, Toutkoushian, & Shah, 2019). Item difficulty is graphed in relation to LT level

with the expectation of increasing difficulty as levels increase (Figures 5a, 5b). Our validation process distinguishes three types of variation in item difficulty:

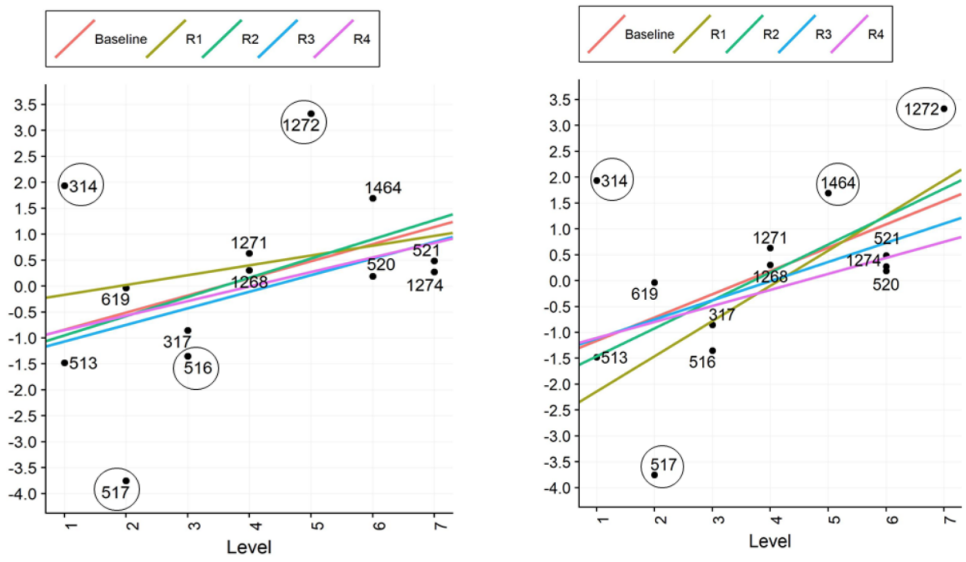
- Inter-level variation: we expect average item difficulty to increase with level.
- Intra-level variation: we maintain variation within a level in order to ensure students have an opportunity to fully explore the breadth of levels in a construct.
- Construct-irrelevant variation: we seek to minimize this source of variation.

To identify items that potentially fail to conform to their assigned level, we conduct a sequential regression analysis. We remove the item with the highest absolute residual value (the most non-conforming) and conduct subsequent regressions until the adjusted R^2 value stops improving, or we have removed 30% of the items. However, unlike conventional psychometric procedures, these items are not simply discarded. The learning sciences (LS) team examines the non-conforming items and recommends one of the following actions:

1. Remove item for lack of alignment to the LT level or excessive construct-irrelevant properties;
2. Relocate item to another level;
3. Adjust item to potentially achieve closer alignment to the central tendency of the difficulty of the level;
4. Flag item as easy/difficult at the level in order to highlight its instructional usefulness to teachers despite its atypical difficulty parameter;
5. Retain the item as is until more data is collected at the precise grade level or grade band before any further action is taken;
6. Identify a set of items at a level for relocating to different level, or
7. Identify substantial portions of the LT or the cluster for restructuring, after which the items are subjected to steps 1-5 or more items are written.

Actions 1, 2, 6 or 7 triggers a subsequent mini-validation cycle. For this cluster, four cycles were conducted, yielding four versions of the sequential regressions. The final two versions of sequential regressions for the Qualitative Graphing construct are shown (Figures 5a, 5b). Each of the sequential regression lines is color-coded per the key; the five items identified as potentially non-conforming are circled.

When the LS team reviewed the circled items in Figure 5a, we conjectured that the variation in difficulty of two of the items at L1 (314 and 513) was due to a compelling misconception that the graph of position vs. time mirrored the physical motion. In item 314 (Figure 1a), 35% of student responses describe the person as walking downhill. Item 513 did not address the misconception and yielded a lower difficulty parameter, suggesting the misconception may be making item 314 atypically difficult for L1. Item 314 was revised to remove a second part that addressed the same misconception but in the context of "going to the store". The revised item was retained to encourage teachers to address that misconception early in the trajectory.



Figures 5a, 5b. Versions 3 and 4 of the Sequential Regression for the Qualitative Graphing LT.

Item 517, which required students to match graphs representing bacteria growth in a petri dish to verbal descriptions, we conjectured was too easy because the number of graphs and verbal descriptions were equal. The item could be answered using 'test-taking' skills without having an understanding of the mathematical content. We revised the item to have one less description. Item 516 (Figure 1b) was revised to include an additional answer option (D) and was no longer flagged in the fourth version due to the re-ordering of levels.

Item 1272 provoked a further examination and revision of the LT, specifically in the order of the top three levels (Table 1). The psychometric model estimates and the data from students across various classes and schools indicated a problem with the difficulty of Level 5 relative to that of Levels 6 and 7. Examining all three levels and related items, the LS team realised that as the LT shifted to cases of functions involving curves, the description of the levels had unintentionally shifted to a focus on rate. (For example, L6 in the original LT read "Interprets continuous changes in rate for basic curve shapes in multiple contexts".) Further, an examination of the problems at these levels revealed that they actually focused on position and time with rate as a secondary consideration (see the item in Figure 1d).

At the same time, the LS team realised that while the items switched between asking students to interpret graphs and sketch curves, the level descriptions lacked this bidirectionality. Therefore, the team revised level descriptions to begin with "Qualitatively interprets or graphically represents..." to convey the bidirectionality to teachers. By reconceptualising this use of an introductory phrase, it became more apparent that the levels should address a clear sequence of cases all the way up the LT: a single linear graph, piecewise linear graphs, simple curves, and complex curves. Items and levels were revised to reflect this new sequence; the level relating position-time and velocity-time graphs was relocated to the top of the LT.

Relocating these levels triggered our fourth and final mini-validation cycle. When the LS team reviewed the circled items as shown in Figure 5b, item 1464 (Figure 1c) popped out as a newly circled, potentially non-conforming item. 73% of the incorrect responses involved an upside-down 'u' or 'v' shape which may resemble, to the students, the movement of the sun across the sky. These results once again suggest the difficulty of the item may be inflated due to a lack of opportunity to encounter and learn from such a misconception. Item 1464 was flagged as difficult but retained in hopes that it would provoke

teachers to address this in their initial instruction and students to think more deeply about their own experiences with shadow length.

In this discussion, all of the interactions around the trading zone combined in a validation process that involved, at least implicitly, all parties. It began with student performance and assumptions about the measurement of LTs. It relied on the LTs' foundation in research on learning, in general consideration of such issues as multiple representations and misconceptions, but remained firmly grounded in research on qualitative graphing as interpreted by the LS team. The psychometric data were, however, interpreted in relation to the particular relevant standards, likely treatment in curriculum, and instruction relative to typical classroom practice. By studying the return of data and how topics are treated, the researchers can also take into consideration the expertise of their partners in the classroom.

Conclusions

Learning trajectories can function as a valuable foundation for instruction. They can be used to synthesise research across a variety of studies on learning and can be written at an appropriate grain size to provide teachers with actionable information on empirically established patterns of student thinking along the way to understanding big ideas in mathematics. We have suggested that students and teachers would benefit from diagnostic classroom assessments that provide evidence on students' progress along such LTs.

Our entire research agenda represents an effort to bring research on student learning and LTs together to influence instructional practice towards learner-centeredness and improvement of student learning. To realise such a vision, we built Math-Mapper, a software tool designed to provide feedback in real time by way of LT-based diagnostic assessments.

We have drawn on Lehrer's (2013) use of the trading zone metaphor to describe the production of LTs for informing and improving classroom instruction, and explored its usefulness to our initiative working at the scale of the entire middle grade curriculum, across multiple sites. We described here the evolution of one LT/LP in MM, and sought to illustrate how it involved deep collaborations among learning scientists, psychometricians, and practitioners. The trading zone is active, dynamic, and multi-dimensional, each party with its own professional community and standards of practice. At the same time, each party must respect the other communities involved.

For Qualitative Graphing, learning scientists recognise the important complement it provides to students' development of formal knowledge of relations and functions. They contribute their understanding of levels of sophistication in student reasoning about (informally) relating graphs to situations of movement or change.

The psychometric perspective brings other expertise and standard practices to the table, many forged in high-stakes tests, and which are challenged by the proximity of classroom assessments to instruction. Discovering new standards of practice in this context of classroom assessment is an ongoing process (Wilson, 2018) and a critical new path for the psychometrics community. For instance, most psychometricians are accustomed to consulting with content experts early in the test development and validation process. However, the theoretical and conceptual requirements of LTs necessitate ongoing collaboration with content experts in order to ensure that the assessments capture and provide feedback on learning that mirrors research on how students learn. Through the development of items and item descriptions, the design of tests, and the review and validation of assessments, psychometricians and learning scientists work together to build and strengthen the measures and the features of MM. This paper has illustrated how, by incorporating expert knowledge on learning theories and embracing the realities of classroom contexts, the trading zone enables psychometricians to shift from providing more passive evaluations of students to taking an active role in understanding and supporting student learning.

Practitioners' roles and the effects of practice on the trading zone were explored in relation to multiple issues. We described the coordination of standards and LT levels and explained the distinctive roles of each. We also focused on the LTs as they function within teachers' data review and illustrated the lively process of bringing students, as active partners, into the assessment-for-learning process. We discussed contributions of practitioners to the process of revising and refining the LTs to fit the conditions of practice, and emphasized the need for observations of practice to inform data interpretation.

Taking LTs to scale, we suggest, puts considerable pressure on the meaning of the trading zone when large numbers of practitioners are involved. Implementation involves a complex system with multiple types of practitioners, from teachers to supervisors to administrators, and they operate within their own system of interactions. While we reported on a number of rich interactions and contributions to features by practitioners in the trading zone, and especially acknowledged the ongoing contributions of a small number of trusted partners, we recognised that while it is essential to collect continuous feedback from as many participants as possible, it is neither feasible nor advisable to give equal weight to all opinions. We found that this is especially the case as we came to realise the need for a learning trajectory process to gradually increase the teachers' own proficiency and understanding of the application.

In sum, envisioning learning trajectories as the product of a trading zone requires significant adaptations by all parties. It requires brokering of the differences in opportunities and constraints. With today's digital resources, and with frequent changes (in knowledge, tools, standards, curriculum materials, instructional systems, and research on learning in diverse communities) that will affect LT/LPs, it makes good sense to build a dynamic system informed by collaborating experts. Only then will we be able to fully and equitably realise the potential of the powerful transformative idea of LT/LPs in education.

References

Adams, R. J., Jackson, J., & Turner, R. (2018). *Learning progressions as an inclusive solution to global education monitoring*. Melbourne, Australia: Australian Council for Educational Research. https://research.acer.edu.au/monitoring_learning/32

Battista, M. T. (2011). Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. *The Mathematics Enthusiast*, *8*(3), 507-570.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. T. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 127-162). New York, NY: Springer.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, *6*(2), 81-89.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812-850.

Common Core State Standards Initiative. (2010). *Common Core State Standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf.

Confrey, J. (2015). Some possible implications of data-intensive research in education—The value of learning maps and evidence-centered design of assessment to educational data mining. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 79-87). Washington, DC: Computing Research Association.

Confrey, J. (2019). A synthesis of research of learning trajectories/progressions in mathematics. *OECD Education 2030 Program.* http://www.oecd.org/education/2030/A-Synthesis-of-Research-on-Learning-Trajectories-Progressions-in-Mathematics.pdf

Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *ZDM Mathematics Education 49*(5), 717-734.

Confrey, J., Maloney, A., Belcher, M. P., McGowan, W. P., Hennessey, M. P., & Shah, M. (2018). The concept of an agile curriculum as applied to a middle school mathematics digital learning system (DLS), *International Journal of Educational Research*, *92*, 158-172.

Confrey, J., Maloney, A., & Corley, A. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM Mathematics Education, 46*(5), 719-733.

Confrey, J., Maloney, A. P., & Nguyen, K. H. (2014). Learning trajectories in mathematics. Introduction. In A. P. Maloney, J. Confrey, & K. H. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. xi-xxii). Charlotte: Information Age.

Confrey, J. & Toutkoushian, E. (in press). A Validation approach to middle-grades learning trajectories within a digital learning system applied to the "Measurement of characteristics of circles". In J. Bostic, E. Krupa, and J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives*. New York, NY: Routledge.

Confrey, J., Toutkoushian, E. P., & Shah, M. (2019). A validation argument from soup to nuts: Assessing progress on learning trajectories for middle school mathematics. *Applied Measurement in Education, 32*(1), 23-42.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Philadelphia, PA: Consortium for Policy Research in Education.

Daro, P., Mosher, F.A., & Corcoran, T. (2011). *Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction* (Research Report No.68). Madison, WI: Consortium for Policy Research in Education.

de Beer, H., Gravemeijer, K., & van Eijck, M. (2015). Discrete and continuous reasoning about change in primary school classrooms. *ZDM Mathematics Education, 47*(6), 981-996.

Dysarz, K. (2018). "Checking In: Are math assignments measuring up?" (Washington, DC: The Education Trust) https://edtrust.org/resource/checking-in-are-math-assignments-measuring-up/

Gonski, D., Arcus, T., Boston, K., Gould, V., Johnson, W., O'Brien, L., . . . Roberts, M. (2018). *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools.* Canberra: Commonwealth of Australia.

Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning, 1*(2), 155-177.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Washington, DC: Chief Council of State School Officers.

Kaput, J. J., & Roschelle, J. (2013). The mathematics of change and variation from a millennial perspective: New content, new context. In *The SimCalc Vision and Contributions* (pp. 13-26). Springer, Dordrecht.

Lehrer, R. (2013). A learning progression emerges in a trading zone of professional community and identity. In R. L. Mayes & L. L. Hatfield (Eds.), *Quantitative reasoning in mathematics and science education: Papers from an international STEM research symposium monograph no. 3*, (pp. 173-186). Laramie, WY: University of Wyoming.

Lehrer, R., Jones, R. S., Pfaff, E., & Shinohara, M. (2017). *Data modeling supports the development of statistical reasoning*. Final report submitted to the Institute of Education Sciences. Washington, DC: U.S. Department of Education.

McCallum, W. (2015). The common core state standards in mathematics. In S. J. Cho (Ed.), *Selected regular lectures from the 12th international congress on mathematical education* (pp. 547-560). Cham: Springer International Publishing.

Monk, S. (1992). Students' understanding of a function given by a physical model. In G. Harel & E. Dubinsky (Eds.), *The concept of function: Aspects of epistemology and pedagogy, MAA Notes, Vol. 25* (pp. 175--194). Washington, DC: Mathematical Association of America.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press. http://www.nextgenscience.org/next-generation-science-standards

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas.* Basic Books, Inc.

Pham, D., Bauer, M., Wylie, C., & Wells, C. (2017, October). *Using cognitive diagnosis models to evaluate a learning progression theory*. Trumbull, CT: Northeastern Educational Research Association.

Piaget, J. (1970). *Genetic epistemology*. New York, NY: W.W. Norton & Company, Inc.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59-81.

Petit, M. M., Laird, R. E., Marsden, E. L., & Ebby, C. B. (2015). *A focus on fractions: Bringing research to the classroom*. New York, NY: Routledge.

Roschelle, J., Kaput, J., & Stroup, W. (2000). SimCalc: Accelerating student engagement with the mathematics of change. In M.J. Jacobsen & R.B. Kozma (Eds.), *Learning the sciences of the 21st century: Research, design, and implementing advanced technology learning environments* (pp.47-75). Hillsdale, NJ: Earlbaum.

Siemon, D., Callingham, R., Day, L., Horne, M., Seah, R., Stephens, M., & Watson, J. (2018). From research to practice: The case of mathematical reasoning. In *Making waves, opening spaces (Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia)* (pp. 40-49). Mathematics Education Research Group of Australasia Inc.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*, 114-145.

Stacey, K., & Steinle, V. (2006). A case of the inapplicability of the Rasch model: Mapping conceptual learning. *Mathematics Education Research Journal, 18*(2), 77-92.

Suh, J., &, Seshaiyer, P. (2015). Examining teachers' understanding of the mathematical learning progression through vertical articulation during lesson study. *Journal of Mathematics Teacher Education 18*(3), 207–229.

Supovitz, J., Ebby, C., & Sirinides, P. (2013). *Teacher analysis of student knowledge (TASK): A measure of learning trajectory-oriented formative assessment*. Philadelphia, PA: Consortium for Policy Research in Education.

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Psicología Educativa*, *20*(2), 109-115.

Vinner, S. (1983). Concept definition, concept image and the notion of function. *International Journal of Mathematical Education in Science and Technology, 14*(3), 293-305.

Wilhelm, J. A., & Confrey, J. (2003). Projecting rate of change in the context of motion onto the context of money. *International Journal of Mathematical Education in Science and Technology*, *34*(6), 887-904.

Wilson, M. (2012). Responding to the challenge that learning progressions pose to measurement practice: Hypothesised links between dimensions of the outcome progression. In A. C. Alonzo & A. Wenk Gotwals (Eds.), *Learning progressions in science* (pp. 317-344). Boston, MA: Sense Publishers.

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*(9), 3766-3774.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, *37*(1), 5-20.

Wilson, P. H., Mojica, G. F., & Confrey, J. (2013). Learning trajectories in teacher education: Supporting teachers' understandings of students' mathematical thinking. *The Journal of Mathematical Behavior*, *32*(2), 103-121.

Wilson, P.H., Sztajn, P., Edgington, C., & Confrey, J. (2014). Teachers' use of their mathematical knowledge for teaching in learning a mathematics learning trajectory. *Journal of Mathematics Teacher Education, 15*(2),149-175.