THE CONCEPTUAL AND PRACTICAL CHALLENGES OF TAKING LEARNING TRAJECTORIES TO SCALE IN MIDDLE SCHOOL MATH

Jere Confrey
NC State University
jconfre@ncsu.edu

Meetal Shah
NC State University
mshah2@ncsu.edu

Jennifer Persson
NC State University
jjvennes@ncsu.edu

Dagmara Ciliano
NC State University
dcilian@ncsu.edu

This paper reports on a design-based implementation study of the use of a diagnostic classroom assessment tool framed on learning trajectories (LTs) for middle grades mathematics, where teachers and students are provided immediate data on students' progress along LTs. The study answers the question: "How can one characterize the challenges encountered when a school implements a diagnostic assessment system around learning trajectories at scale?" by identifying three explanatory themes: shifting to classroom assessment, understanding the concept and content of the LT, and seeing the results as a call to action. Each theme is discussed with references to observed activities and discussions with participants and related to the challenges connected with taking the concept of LTs to scale.

Keywords: Learning Trajectories, Assessment and Evaluation, Teacher Knowledge

Introduction

Many believe that LTs hold great promise for widely strengthening mathematics instruction by informing teachers about the knowledge of the empirical patterns on how students learn (Daro, Mosher, & Corcoran, 2011). However, locating the disparate research contributions poses a significant risk to influencing practice at scale. Confrey and colleagues have sought to address this need by creating a software tool, Math Mapper 6-8 (MM)(sudds.co), organized around a learning map of nine big ideas, 25 relational learning clusters (RLCs) and 62 constructs. Each construct delineates a LT based on a synthesis of the related research (Confrey, 2015) that draws from the same research base as turnonccmath.net (Confrey & Maloney, 2012). MM is based on the idea of a LT as a research-based model of how students' thinking increases in sophistication relative to a domain-specific concept, in the context of instruction that is operationalized through the use of digitally-administered and scored diagnostic assessments, which return data to students and teachers immediately. These 30-minute assessments consist of items that are aligned with the levels of the LTs and to avoid excessive testing, include the content covered in an RLC. Because LTs can span multiple grades, teachers can select relevant grade-level tests (6, 6-7, 7, 7-8, 8, 6-8). Multiple equivalent forms of a test are administered in a classroom to ensure all LT-levels are assessed across students[1]. Assessment items are written by the team in consultation with inservice teachers and designed to elicit student thinking and raise issues worthy of classroom discussion. We have reported elsewhere on the validation (using item response theory (IRT)) of the trajectories based on data from students with varied demographics from our six partner middle schools (Confrey & Toutkoushian, 2018; Confrey, Toutkoushian, Shah, 2019). Our goal is to use the tool at scale across all teachers and all topics (except Algebra 1) in middle school(s) to strengthen instruction based on empirical results from our assessments.

We would argue that mathematics education needs to tackle more issues at scale and specifically how to improve learning for all students as evidenced on valid, reliable, and equitable measures.

Study after study has demonstrated the naivety of assuming that data from assessments alone sufficiently informs instruction that leads to more learning (Nelson, Slavit, & Deuel, 2012). Likewise, our studies of the use of MM reinforce the view that implementation of multifaceted learning systems is a complex activity requiring significant professional support and attention to organizational factors (Mandinach, Gummer, Muller, 2011; Tyack and Cuban, 1995). Recognizing this complexity, we asked the research question: “How can one characterize the challenges encountered when a school implements a diagnostic assessment system around learning trajectories at scale?” We recognized that the answer to this question would have conceptual and practical components.

Introducing new forms of classroom assessment in schools frequently has to overcome the barriers formed by negative reactions to the high-stakes testing required by *No Child Left Behind*. However, there remains an appetite for formative assessment practices (Black & Wiliam, 1998; Brookhart, 2015; Heritage, 2008). Classroom assessment is designed to focus on student learning and growth, rather than to view assessment as a means to measure summative accomplishment (Heritage, 2008; Wilson, 2018). Heritage describes these as including a clear statement of the learning goal, an emphasis on self-regulated learning, and focus on movement along a learning progression. She, and others, emphasize the use of assessment *for* learning (Black, Harrison, Lee, Marshall, & Wiliam, 2004) and stress that using an assessment formatively depends not on the instrument, but *how* it is used practically. It requires the focus to be on what the data show about the state of one’s understanding and how to move forward. Such assessment requires students and teachers to shift to a growth mindset (Dweck, 2006).

Math Mapper 6 - 8: A Diagnostic Classroom Assessment Tool

MM is designed to be compatible with varied curricula and scope and sequence documents. Prior to implementation of MM at a site, a few lead teachers align the assessments to relevant timepoints in the school’s scope and sequence. Any teacher can administer any assessment at any time, but a coordinated schedule of assessment supports teacher discussion, analysis, and planning at grade level. Our assessment approach involves having teachers conduct initial instruction (with or without pretesting), and about ⅔ of the way through the allotted instructional time, to give a diagnostic assessment on the relevant material using MM. After testing, teachers initiate data reviews and address topics needing further development as shown in Figure 1.

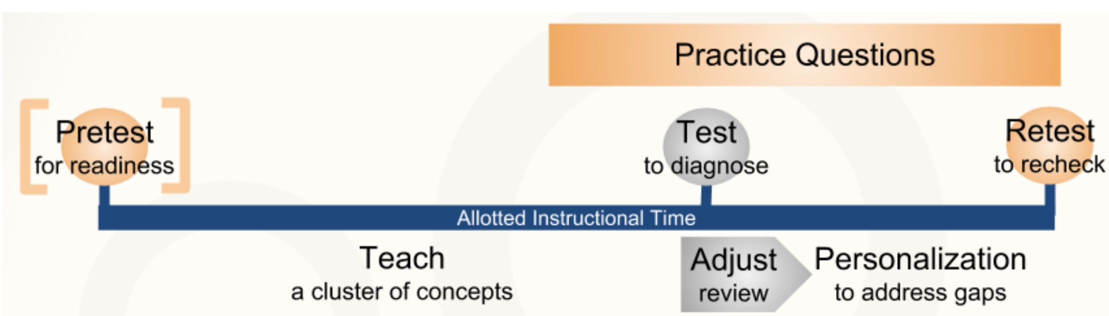

Figure 1: A Model for Implementing Classroom Assessment and Data Review

The student data are returned using a visualization of the cluster from the map with dials reporting the percent correct for each construct. Students can access a LT ladder showing the levels tested and their score by level. Students can also scroll to an item matrix showing items by level where they can review their responses and revise and resubmit answers. The teacher's display is called a heat map (Figure 2) where she sees the student performances ordered in columns from weakest to strongest[2] by construct and the levels ordered from lowest to highest in rows. The white boxes indicate untested levels for the student whose data are in that column. The other boxes are colored coded from orange (incorrect) through shades of blue to darker blue (correct). Teachers are taught to approximate Guttman[3] curves in order to identify which levels need re-teaching and which students need additional help.

The teachers have routinized two approaches to data return. In the first, using whole class instruction, they decide which levels to review based on the heat map. They tend to look for a level that is predominantly orange. Then they open it to view the item. The item can be viewed with or without the correct answer, an item analysis of student responses, and/or a report on the frequency of common misconceptions. Teachers vary substantially in the degree of student involvement in the review process, despite the research team's efforts to promote learner-centered reviews. The second approach developed by teachers is to use the data to form student groups (usually homogeneously exhibiting similar error patterns) to discuss the problems, revise, and resubmit. There is a practice feature in MM at the construct level, where individuals or groups of students can access additional items at levels of their choice and receive immediate feedback on the correctness of their responses.

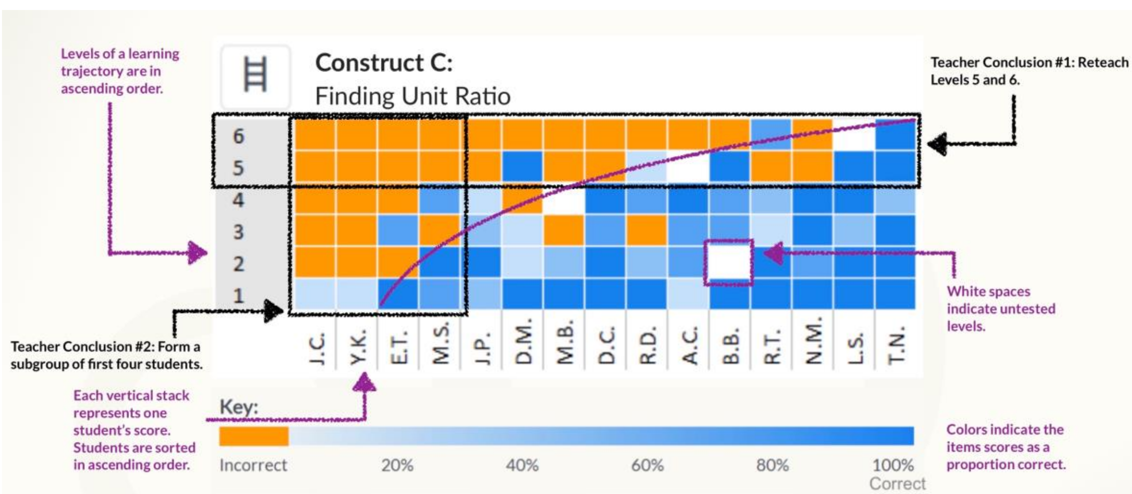

Figure 2: Sample Heat Map on "Defining and Measuring Center" With Labeled Components

Theory

The theoretical approach to the study is grounded in constructivism (Confrey and Kazak, 2006; Steffe & Gale, 1995; von Glasersfeld, 1982), with its focus on understanding how students build their knowledge gradually, working through carefully sequenced tasks in the company of peers, building gradual understanding. The process exemplifies what Piaget called "genetic epistemology" (Piaget, 1970) and Freudenthal and colleagues called "guided reinvention"

(Freudental, 1991; Gravemeijer & Doorman, 1999). It rests on the central construct of a learning trajectory (Clements and Sarama, 2004; Confrey, 2019; Simon, 1995) as a

> researcher-conjectured, empirically supported description of the ordered network of constructs a student encounters through instruction (i.e., activities, tasks, tools, forms of interaction and methods of evaluation), in order to move from informal ideas, through successive refinements of representation, articulation, and reflection, towards increasingly complex concepts over time. (Confrey et al, 2009, p. 347)

Secondarily, the research draws on a socio-cultural perspective on how students in classrooms engage in mathematical practice, share their approaches, cultural experiences, resources and insights, and gradually internalize the mathematical norms and expectations of a field in classroom practices (Lehrer and Schauble, 2000; Vygotsky, 1978; Yackel & Cobb, 1996). The project also draws extensively from the research on teachers' professional knowledge of content, pedagogical content knowledge (Shulman, 1986) and mathematical knowledge for teaching (Ball, Thames, Phelps, 2008), including LT-based instruction (LTBI)(Sztajn, Confrey, Wilson, & Edgington, 2012). For teachers to successfully achieve a learner-centered classroom (Confrey et al., 2017) engaging students in highly productive tasks (Stein, Engle, Smith & Hughes, 2008), they must seek to draw out student ideas and know how to orchestrate successful discussions from emergent models (Gravemeijer, 1999). Finally, our approach draws on the literature on professional growth by teachers sharing and discussing data in professional learning communities (PLCs) (Grossman, Wineburg, & Woolworth, 2001; Mandinach et al., 2011).

Methodology and Data Sources

This methodological work is situated as "design-based implementation research" (DBIR) carried out with our six demographically diverse research partner schools in 3 districts (Fishman, Penuel, Allen, Cheng, Sabelli, 2013). The schools approached the research team, wanting either to implement forms of "classroom assessment" (Pellegrino, Chudowsky, & Glaser, 2001) in which their teachers could receive data in a timely way during instructional units to revise and improve instruction and/or wanting more information about LTs. The collaboration among teachers, learning scientists, psychometricians, and software engineers involved conversations with school leadership and curriculum supervisors, from 2-4 days of summer professional development (PD) on the tool and underlying approach to learning trajectories, and then implementation of the assessments periodically during the year customized to the school's curriculum. Feedback to the research team occurs during regular grade-level PLC meetings where teachers reviewed data, discussed challenges, requested additional features and learned from peers. Data for this study were collected digitally through the use of the assessment system (n = 62000 tests), through observations and video records of classroom data returns, PLC meetings, and PD meetings meeting notes with school leadership, and monitoring ongoing participation in communication networks among teachers and researchers. Analysis of data for this paper was undertaken by the research team reviewing the video and artifacts to understand how teachers implemented the software and interpreted and acted on the data. From the classroom observations, the data review and discussions with district leaders, a set of three themes emerged to describe and explain the challenges inherent in using the software as intended to strengthen learner-centered practices and increase learning. They are summarized and discussed in terms of their conceptual and practical implications. Further, they are offered as hypotheses for future research.

Results

Theme 1: Shifting to Classroom Assessment

The observations of data returns by teachers suggest that in order to change typical views of testing by teachers and students, a shift to classroom assessment requires intentionality, explicit actions, and discussion with students. According to teachers, most students view tests as indicators of how knowledgeable, smart, and hard-working they are, and anticipate the results with trepidation and anxiety. However, classroom assessments, used formatively, are intended for feedback rather than personal evaluation and judgment. The diagnostic feedback should support informed decision-making and actions, the involvement of the students as partners, and the use of student thinking to inform next steps.

Based on our observations and teacher reports, many teachers quickly draw students' attention to their opportunity to revise and resubmit with MM. Students appreciated the opportunity to rework problems and they frequently expressed surprise that simply by rereading the problems and trying harder, they could get correct answers and experience the reward of seeing the dials immediately show their improvement. This is perhaps the most evident, simple, and direct example of the tool being used for classroom assessment.

Teachers who chose to group students together based on the heat maps seemed to be most successful in using the tool to strengthen attention to students' self-regulation, a key element of classroom assessment. One teacher, after organizing his students into groups, requested that they practice in constructs on the levels needing improvement based on their results and then return to revise and resubmit incorrect answers. His goal was to encourage them to learn the level and not just the item. Observations of his groups showed students explaining the ideas about the measurement of circles successfully to each other, calling over the teacher when more help was needed. These examples represent successful transitions to classroom assessment.

Observations also tended to reveal many teachers using the heat maps for data review by pulling up items from predominantly orange levels and simply again telling students how to solve the problems. They included admonitions to students to recall prior advice such as "I have told you to begin by drawing a T chart and building a table". These teachers seemed to view students' weak performance as needing quick and direct remediation rather than as opportunities to examine student thinking. Over time and with encouragement, teachers began to recognize that the students, having worked the problems, could provide valuable insights in their thinking. For instance, one teacher, on review of the data, recognized that she had neglected to teach percents greater than 100. She used an item at this earlier level, where the percent was 200, to reteach the concept and then relied extensively on student contributions to solve an item at a higher level involving 245% (Confrey, Maloney, Belcher, McGowan, Hennessey, & Shah, 2019). She referred to the MM items as "stretch items" and helped her students recognize their own potential to solve them. These observations have led us to recognize that even though the data provide direct evidence of students' learning needs, many teachers, especially in our lower performing schools, need additional support to learn how to orient their instruction to actively draw on students' thinking and utilize tenets of productive discourse (Stein et al., 2008).

A major challenge in shifting the orientation to a growth mindset occurred due to the weakness in overall performance by students, which may be due to the assessment's focus on conceptual understanding and reasoning. The score averages by cluster typically range from 40-60% correct, and are thus approximately 20-30% lower than on typical unit tests and quizzes. For students to understand these lower scores, teachers need to help students understand that these assessments are diagnostic and that in order to provide valuable information for all students, are designed to result in lower scores. This lower range is essential to allow for space in which to measure growth. Even so, the research team has also been concerned with the extent of the weakness in student performance, and has checked the alignment with grade level standards and asked students to judge if the material has been taught; they confirm it has. It is possible that the weak performance is an indicator of excessive procedural instruction. This would be consistent with other research which reports that middle grades students are being given excessive amounts of procedure-based materials (Dysarz, 2018) and that many teachers struggle to distinguish procedural understanding from higher conceptual levels, much less, LT-based levels (Supovitz, Ebby, and Sirinides, 2013). Further evidence for this emerged from some schools within the other themes where we discuss its implications for our future work.

How teachers responded and handled the challenge of shifting to classroom assessment and focusing on learning varied significantly by school. Schools with strong internal professional community supervisors, coaching, and district leadership transitioned more easily. In those settings, the teachers mediated the student responses, helping students to see low percentages simply meant "there was more work to do." She encouraged students to persevere by saying "our average was at 70% but this was just the first time… you'll have a chance to revise", and later, after students had revised much of their work, saying "If you're a risk-taker you can try a higher level." The strongest teachers focused on the content of the items, drawing connections to similar or related problems they had done, how to work through their reasoning, and how to coordinate the use of a variety of representations. Others focused on improvement, reporting back that "we have doubled our average score" and on refreshing the heat map to show all the students who had revised and resubmitted correct responses. In settings in which competing initiatives, especially around assessments, provided different data and direction, or mentoring and supervision were absent or weak, the initiatives encountered more problems.

These observations illustrate the complications of moving towards measurement-oriented classroom assessment. It appears that prior expectations influence the interpretation of the scale and that only if teachers explain reasons for the differences, focus on growth and the content itself, do they successfully shift the class's orientation towards using assessment for learning.

Theme 2: Understanding the Concept and Content of the LT

A second challenge of working at scale with MM comes from the need to assist teachers in understanding the conceptual foundations of the LTs in the map. The learning map in MM provides teachers access to all 62 LTs and the related misconceptions. Common Core State Standards are identified and aligned to each construct, and each level is mapped to its projected grade level. During a 2-day PD workshop, teachers are introduced to the conceptualization and research underpinning two clusters on ratio within the big idea of "compare quantities as ratio, rate or percent and operate with them". This consists of discussions of the relationship among the three constructs of ratio equivalence, base ratio, and unit ratio; and of how these form the foundation for building up, comparing ratios, and finding missing values, as well as the sequencing of the levels within each construct.

In going to scale with LTs, we have found that not reviewing all the LTs at the same level of detail and simply providing access to the LTs is insufficient for affecting practice. Observations at PLC meetings indicate that teachers seldom review LTs in planning instruction. The distinctions between levels and sequencing of levels are often overlooked by teachers. We anticipated that this would be the case, but we had hoped that providing the teachers data on students' performance would result in teachers recognizing the LT's value and relevance.

LT-based assessments differ from typical assessments that comprise of items that sample the various topics in a content domain, often referred to as "domain sampled" assessments. When teachers give domain-sampled tests, they often review only the difficult items, sometimes followed with extra explanations or practice. The item is viewed as a case of that portion of the domain. In an LT-based assessment, the item is also a case, but it is a case of likely student reasoning for that level of the construct. The meaning of an item therefore is situated in a construct for that level, and moreover, the level is situated in a sequence which delineates prior and subsequent ideas. When we observe most teachers reviewing data, the item is simply treated as an item to solve. This lack of recognition of the role and value of the trajectory became evident at one of the PLC meetings when a teacher said, the "levels [on the heat map] were random". If teachers do not recognize the significance of the LT, then much of the potential efficiency of the approach is lost.

Some teachers showed difficulty in understanding the structure of the LTs. During one of the PD sessions, the teachers complained that the ratio problems included internal multiplicative relationships that were too difficult, saying "[MM] gets too decimal-y [sic] and into fractions too quickly. We need to stick with 'numbers'." These comments showed a lack of familiarity with the Finding Missing Values in Proportions LT and strategies to find missing values in a 2 by 2 ratio box. Level 1 begins with whole number multipliers within and across the proportions, level 2 moves to combinations of multiplication and division (e.g. multiplication by 3 and division by 20, referred to as daisy chains (Confrey et al., 2014)), and level 3 tests for the resultant rational number operation (e.g. $4 \times \frac{3}{20}$). After reviewing this approach, teachers' positive reactions suggest a lack of familiarity with pedagogical approaches like daisy chains which make multiplication by a rational number more accessible to students.

Some teachers indicated that they expected the assessments items to "mirror" the items that they had taught in class. When reviewing the data, they advised students to solve the item procedurally rather than urge students to explain their thinking or engender discourse around the item. For instance, one teacher stated that, "Any time you are given three values and one unknown, that's kind of a hint that this is proportions". Such an approach is unlikely to support students in recognizing the fundamental multiplicative relations inherent in proportions (levels 1-3) and, subsequently, in distinguishing proportional from non-proportional relations (level 6).

It is becoming increasingly clear that to effectively use the tool, districts and schools will have to invest significantly in PD around the meaning of the LTs. Our experience has convinced us to begin to provide further information about the LTs and how they are situated in clusters. We see significant professional opportunities in also working with others who use other forms of evidence of student progress on LTs such as work samples (Petit, 2011; Suh & Seshaiyer, 2015).

Theme 3: Seeing the Results of LT Assessments as a Call to Action

As a diagnostic assessment tool, MM highlights issues of student understanding, and while the LTs can point out directions for movement, the effectiveness of the tool depends on the actions taken by its users (students and teachers). Responding to MM's results can be particularly challenging because classroom assessments from a robust LT-based diagnostic assessment can initially result in substantially lower student scores than other more traditional or teacher-created tests, especially if these common assessments focus primarily on procedures. We observed contrasting teachers' responses to their students' weaker performance data on these diagnostic assessments. Some teachers approached these results as a challenge, or as a call to action, encouraging their students to revise their work while simultaneously displaying the class's immediately increasing scores in real-time on a screen at the front of the classrooms, as students continued to work on these revisions. Other teachers exhibited resistance to the data.

One type of resistance that emerged was by questioning the test itself: one teacher felt that he should be able to anticipate his students' scores before they take any test, and if the scores are not what he expected, then clearly something was wrong with the test's ability to accurately assess his students' abilities. Secondly, teachers expressed beliefs that this LT-based assessment is not aligned to their curriculum, or is not aligned to "how I taught it". It is important to note that the LTs within MM and these teachers' curriculums are both aligned with Common Core State Standards, which means that these two systems are not misaligned, as some teachers claimed. A third concern of teachers is viewing the data as a form of exposure for them personally, as the results vary considerably from teacher to teacher, even within the same school. There is clear apprehension that the results will be used to evaluate them. Administrators played a key role in how this concern played out. In our highest performing school, an able mathematics supervisor kept discussions focused on the students and how to use the data to meet their needs. In a lower performing district, the administrator emphasized that low scores should not be the focus, but demonstrations of improvement should. In a third setting, with more site-based orientation, the degree of accountability ranged from strong to weak based on the instructional leadership provided by principals and other administrators.

When teachers encountered results that were lower than they expected and responded by only or excessively expressing concerns with the measure itself or of curricular/instructional misalignment, the research team noted that the response allowed them to avoid any sense of accountability for their students' LT-based data. Most often such responses to MM and the data occurred among the same teachers who expressed a preference for the use of highly procedural practices. For instance, in one school, teachers expressed a preference for using a computer system that focuses primarily on procedures. In another, teachers avoid more complex or conceptual orientations by developing their own simplified curricular materials. Thus, these observations suggest that if districts and schools want their teachers to view the results of an LT assessment as a call to action, rather than resist and reject the information, then additional supports need to be put in place to ensure their teachers understand and value an LT-based approach to learning, over a procedural approach.

Conclusions

In this paper we describe a disruptive innovation, MM, (Christensen, Raynor, & McDonald, 2015) that sits at the intersection of classroom formative assessment theory and LTs. We propose a critical goal of taking such an innovation to scale is to strengthen instruction through a model of personalization that is driven by data from valid, reliable, and equitable measures of student learning. However, taking any innovation to scale requires iterative cycles of "ramping up" toward full and successful implementation, informed by insights from classroom practice. Our DBIR study exposed important insights into the fit between MM's design and typical classroom practice. We characterized insights from our classroom observations into three preliminary themes as a means to describe the necessary shifts in practice, the need for teacher supports around LTs, and required collaborations among administrators at schools/districts. We see these descriptions as informing the development of "guardrails" to increase the likelihood of successful implementation at scale of MM. Our study also demonstrates that in order to realize the promise of LTs at scale, more resources must be devoted to helping teachers understand the foundation of each LT and cluster.

Endnotes

1 Most RLCs have 3 constructs averaging 6 levels, resulting in 18 possible levels to be tested in an assessment. With shorter tests averaging 8-10 items, not all levels are tested.

2 Teachers can display student initials to aid their own interpretation which is supported by a student matrix below or hide them for anonymity during classroom projection.

3 The display of Guttman curves and related advice on which levels to re-teach and which groups to form are currently planned for automation.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant 1621254.

References

Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8-21.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice,* 5(1), 7-74.

Brookhart, S. M. (2015). *How to make decisions with different kinds of student assessment data*. ASCD.

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

Christensen, C. M., Raynor, M. E., & McDonald, R. (2015). What is disruptive innovation. *Harvard Business Review*, 93(12), 44-53.

Confrey, J., Maloney, A., Belcher, M. P., McGowan, W. P., Hennessey, M. P., Shah, M. (2019). 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. (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., & Kazak, S. (2006). A thirty-year reflection on constructivism in mathematics education in PME. In A. Guttierez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education: Past, present and future* (pp. 305–345). Rotterdam, the Netherlands: Sense Publishers.

Confrey, J., & Maloney, A. (2012). A next generation digital classroom assessment based on learning trajectories. In C. Dede & J. Richards (Eds.), *Steps toward a digital teaching platform: Customizing classroom learning for each student*, (pp. 134-152). New York, NY: Teachers College Press.

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., Nguyen, K., Mojica, G., & Myers, M. (2009). Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories. In M. Tzekaki, M. Kaldrimidou, & H. Sakonidis (Eds.), *Proceedings of the 33rd conference of the International Group for the Psychology of Mathematics Education* (pp. 345–352). Thessaloniki, Greece: International Group for the Psychology of Mathematics Education.

Confrey, J. & Toutkoushian, E. (2018). 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: Routledge.

Confrey, J., Toutkoushian, E., & 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.

Daro, P., Mosher, F.A., & Corcoran, T. (2011). *Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction*. Consortium for Policy Research in Education Report #RR-68. Philadelphia, PA: Consortium for Policy Research in Education.

Dweck, C. 2006. *Mindset: The new psychology of success.* New York: Random House.

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/

Fishman, B. J., Penuel, W. R., Allen, A. R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National Society for the Study of Education*, *112*(2), 136-156.

Freudenthal, H. (1991). *Revisiting mathematics education: China lectures.* Dordrecht, Netherlands: Kluwer.

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

Gravemeijer, K., & Doorman, M. (1999). Context problems in Realistic Mathematics Education: A calculus course as an example. *Educational Studies in Mathematics, 39*, 111–129.

Grossman, P., Wineburg, S., & Woolworth, S. (2001). Toward a theory of teacher community. *Teachers college record,* 103(6), 942-1012.

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

Lehrer, R., & Schauble, L. (2000). Developing model-based reasoning in mathematics and science. *Journal of Applied Developmental Psychology, 21*(1), 39-48.

Mandinach, E. B., Gummer, E. S., & Muller, R. D. (2011). *The complexities of integrating data-driven decision making into professional preparation in schools of education: It’s harder than you think. Report from an invitational meeting.* Alexandria, VA: CNA Analysis & Solutions.

Nelson, T. H., Slavit, D., & Deuel, A. (2012). Two dimensions of an inquiry stance toward student-learning data. *Teachers College Record*, *114*(8), 1-42

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

Petit, M. (2011). Learning trajectories and adaptive instruction meet the realities of practice. In P. Daro, F. A. Mosher, & T. Corcoran (Eds.), *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (Research Report No. RR-68; pp. 35–40). New York: Consortium for Policy Research in Education.

Piaget, J. (1970). *Genetic epistemology*. trans. E. Duckworth. New York, NY, US: Columbia University Press.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15(*2), 4-14.

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

Steffe, L., & Gale, J. (1995). *Constructivism in education*. Hillsdale, NJ, USA: Lawrence Erlbaum.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical thinking and learning*, 10(4), 313-340.

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. B., & Sirinides, P. (2013). Teacher analysis of student knowledge (TASK): A measure of learning trajectory-oriented formative assessment. Retrieved from http://www.cpre.org/sites/default/files/researchreport/1446_taskreport.pdf

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia*. Harvard University Press.

von Glasersfeld, E. (1982). An interpretation of Piaget's constructivism. *Revue Internationale De Philosophie,* 612-635.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Eds. and trans.). Cambridge, MA: Harvard University Press.

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

Yackel, E., & Cobb, P. (1996). Sociomathematical Norms, Argumentation, and Autonomy in Mathematics. *Journal for Research in Mathematics Education, 27*(4), 458-477.