Community for Advancing Discovery Research in Education

Classroom-Based STEM Assessment: Contemporary Issues and Perspectives

AUGUST 2023

Acknowledgments

This report is a product of the Community for Advancing Discovery Research in Education (CADRE) Committee on Classroom-Based Assessment. CADRE connects STEM education researchers funded by the National Science Foundation's Discovery Research PreK-12 (DRK-12) program to share their methods, findings, results, and products with the research and development community and the greater public, to better inform and impact policy, research and education.

Steering Committee:

Christopher J. Harris,
WestEd

Eric Wiebe,
North Carolina State University

Shuchi Grover,
Looking Glass Ventures

James W. Pellegrino,
University of Illinois Chicago

Committee Members:

Eric Banilower,
Horizon Research, Inc.

Arthur Baroody,
University of Illinois at Urbana-Champaign

Erin Furtak,
University of Colorado at Boulder

Ryan "Seth" Jones,
Middle Tennessee State University

Leanne R. Ketterlin-Geller,
Southern Methodist University

Okhee Lee,
New York University

Xiaoming Zhai,
University of Georgia

Suggested Citation:

Harris, C.J. Wiebe, E., Grover, S., & Pellegrino, J.W. (Eds.) (2023). *Classroom-Based STEM assessment: Contemporary issues and perspectives*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

We thank the following individuals for their review of this report: Margaret Heritage, University of California, Los Angeles; Maria Ruiz-Primo, Stanford University; and Amelia Gotwals, Michigan State University

The activity was funded by the National Science Foundation, Grant # 2100823. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Classroom Assessment in STEM Education: An Introduction to the Report

James W. Pellegrino, Shuchi Grover,
Christopher J. Harris, & Eric Wiebe

Report Goals and Context

This report is focused on what research and practice indicate about how assessment impacts the teaching and learning of STEM subject matter in K–12 classrooms. Our intent is to take stock of what we currently know as well as what we need to know to make classroom assessment in STEM maximally beneficial for the instructional practices of teachers and the learning outcomes of students. The volume draws inspiration from the cumulative body of research on STEM education that has accrued over the last two decades, with particular emphasis on work funded by the National Science Foundation (NSF) through its Discovery Research PreK-12 (DRK-12) funding program. As articulated in NSF's DRK-12 solicitation: "For assessment to be a driving knowledge engine that moves STEM education forward it must be integrated with systems of learning and teaching, with specific attention paid to the needs of practitioner communities and how assessments would be used in formal education settings" (National Science Foundation, 2020). The major sections in this report, individually and collectively, focus on critical issues regarding assessment integration and use at the classroom level.

This work is sponsored by the Community for Advancing Discovery Research in Education (CADRE), an NSF-funded network for STEM education researchers endeavoring to improve education in science, technology, engineering, and mathematics (STEM) through various information-sharing and community-building mechanisms (**https://cadrek12.org)**. CADRE's leadership believes that consideration of critical issues in research and development on classroom-based STEM assessment is much-needed and long overdue. Multiple research and policy developments over the last two decades support the potential value of such an effort. Included among them are the following: (a) the changing landscape of policy and practice discussions regarding desired outcomes from the educational system for the 21st century, (b) related changes in the content standards and student outcomes expected in the STEM disciplines, (c) explications of the science underlying design of educational assessments to support their varying forms and functions in the educational system, (d) evolution of theory and research on the nature and development of STEM disciplinary learning and its implications for classroom instruction and assessment, and (e) a shift from the emphasis on large-scale standardized testing to greater focus on the uses of assessment in the classroom as part of ongoing teaching and learning.

Two important societal-level factors make the discussion of classroom instruction and assessment particularly pertinent. The first is increased urgency to address long-standing disparities in opportunities to learn and educational outcomes for underserved and marginalized populations and the role of assessment in reducing differences in opportunities and outcomes rather than reifying them. The second is rapid growth of student access to technologies and the development of potent computational tools, such as data analytics and artificial intelligence (AI), to support the integration of classroom assessment into instructional practices, including enactment of formative assessment practices.

Major Developments Impacting STEM Education and Classroom Assessment

Six major developments reflecting the trends noted above, and which significantly impact conceptualizations of STEM education and classroom assessment, are briefly highlighted below. They establish a context for the material reviewed in and across this report's five major sections. Each development is briefly stated as an assertion. Relevant background information constituting warrants for each assertion is presented subsequently in this introduction section.

1. **Shifts in how we conceptualize standards** and expectations for STEM Proficiency have changed substantially and now demand integrated knowledge of core disciplinary knowledge and practices. The new standards have major implications for design and implementation of curriculum, instruction, and assessment.

2. **Theories, models, and data** on disciplinary knowing and learning have changed substantially and are best represented by a broad sociocultural perspective on the general nature of knowing and learning. These perspectives compel us to re-center STEM learning experiences on students' interests and identities and frame assessment in contexts that are relevant to students' lives, with particular attention to students who have been historically marginalized in school settings.

3. **Coordination and integration of curriculum**, instruction, and assessment is essential to achieve coherent classroom learning environments and best achieved when all three are derived from conceptual models and empirical data on disciplinary knowing and learning, such as discipline-specific learning progressions.

4. **Assessment is a process of reasoning from evidence** represented by the three interconnected elements of cognition, observation, and interpretation. Critical to the validity of this "reasoning from evidence" process are conceptual models and empirical data on disciplinary knowing and learning.

5. **Classroom practice requires differentiating formative from summative functions** of assessment and the ability to implement these practices accordingly. Formative assessment can significantly impact student learning outcomes, but doing so hinges on multiple facets of teachers' assessment literacy knowledge, as well as the availability of tools that teachers can use to support valid and appropriate formative and summative practices.

6. **The affordances of technology to support classroom assessment** have greatly increased, including interactive and adaptive stimulus materials, response data capture, and application of data analytic and computational interpretive tools. Considerable work remains to be done to effectively integrate these developments into classroom instructional practice.

Implications for the Contents of this Report

More than 20 years ago, Shepard (2000) called attention to ways in which classroom assessment practices needed to change to better support student learning. She pointed out that (a) the design of assessments needed to be improved to reflect contemporary views of learning, (b) the gathering and use of assessment information and insights needed to be made integral to the learning process, and (c) assessment must be addressed in teacher preparation programs and in the professional learning of practicing teachers (Shepard, 2000). Her comments at that time were reflective of a growing stance among scholars that if assessment, curriculum, and instruction were better connected so that they work together, student learning would improve. In the interim, we have learned much that bears upon this vision of productive assessment in the STEM classroom as represented in part by the six developments highlighted above.

In attempting to consider what has been learned, our challenge as editors was deciding on focal areas for this report. Decision-making was guided by the substantial bodies of work that have accrued on critical issues in the integration of assessment into classroom teaching and learning in the STEM disciplines *and* that have major implications for current practice as well as future research and development. What follows are brief summaries of the five sections in this report that meet these criteria. Each section's discussion of issues should prove highly productive for stimulating dialogue and future work, in part because it integrates and provides the authors' perspectives regarding knowledge across several of the developments noted above. Collectively, the five sections cover all six developments listed above while representing different and important integrations and perspectives on the knowledge and practice of STEM classroom assessment.

Section 1. Connecting Classroom Assessment with Learning Goals and Instruction through Theories of Learning emphasizes the importance of aligning classroom assessments with STEM learning goals and instruction so that they can be used as part of ongoing instruction to monitor students' progress in learning and to support students in future learning. The authors put forth the position that models of how learning progresses in the STEM disciplines, often referred to as either *learning progressions* or *learning trajectories*, offer a compelling and principled way for developing assessments that align with learning goals and instruction and cohere with learning theory. This alignment is critically important for ensuring that assessments reflect contemporary views on learning in STEM disciplines and that they serve to provide insight into how students' disciplinary knowledge and practices are developing over time with appropriate instruction. The authors discuss how student response data from these kinds of assessments can potentially be used to inform instructional decision-making and improve the teaching and learning process. **Section 1** also describes several NSF DRK-12 projects that are using learning progressions to guide the design and use of classroom assessments. The section concludes with three recommendations for future research and development of classroom-based STEM assessments.

Section 2. Assessment for Learning reviews the conceptual and empirical literature on formative assessment—what it is, why it matters, and evidence for its efficacy—and connects this classroom assessment practice to research on learning progressions or learning trajectories. It illustrates what is possible in the STEM classroom, with detailed illustrations of assessment connected to the early learning of mathematics, while simultaneously highlighting the challenges of doing so. Drawing upon NSF DRK-12-funded projects, it provides examples of designing classroom assessment closely aligned with learning trajectories for important aspects in the development of early mathematical knowledge and skill. The section reviews evidence in support of the impact and efficacy of formative assessment on student learning, including concerns in the literature on evaluation of outcomes. It also highlights some of the major conceptual and empirical work that remains to be done to enable teachers to engage productively in the assessment for learning process as part of their overall classroom instructional practice and includes suggestions for future research and development.

Section 3. Equity and Justice in Classroom Assessment of STEM Learning examines the conceptions of equity and justice in the emerging literature in STEM learning and considers how these contemporary perspectives can be centered in STEM classroom assessment. The authors recognize assessment as one of three central components of a classroom learning environment—curriculum, instruction, and assessment—and argue that centering equity and justice is a transformative act that will require accompanying transformations in curriculum and instruction. Since all three interact within a classroom to shape STEM learning, all three must be transformed for assessment to fully support teaching and learning. The section presents descriptions of several DRK-12 projects to illustrate how equity and justice are addressed in classroom assessments of STEM learning. While many projects focus on equity in assessment, projects that center justice are only beginning to emerge. Recommendations for future directions for equity and justice in STEM classroom assessment are made.

Section 4. Teacher Knowledge and Practices for Assessment. While there is a growing knowledge base on supporting student learning through formative assessment, ultimately, success hinges on whether teachers understand the formative use of assessment and can incorporate it into their classroom practice. Formative assessment is thus a fundamental teaching practice that must become part of preservice and in-service teacher assessment literacy and professional development. In this section, the authors draw on over two decades of research on teacher assessment literacy and practice to argue that teachers need support in three domains to carry out high-quality assessment in their classrooms. First, they need models of student learning that are empirically grounded, developmental, and interpretable for classroom application. Second, teachers need feasible strategies for eliciting valid evidence of student thinking. Third, teachers need tools that support interpretations of the generated evidence in terms of the learning model. The section also makes the argument that learning progressions or learning trajectories are a key vehicle to support teachers in engaging in high-quality classroom assessment practices. Two illustrative examples from NSF's portfolio help bring alive these ideas on teacher knowledge and preparation for classroom assessment: (1) a learning progression to support the development of statistical reasoning and (2) a professional development model for improving chemistry teachers' formative assessment practices. The section ends by identifying challenges and providing recommendations for future directions in STEM teacher preparation, including the unique challenges presented by new subjects, such as computer science and engineering.

Section 5. Technology-Based Innovative Assessment. A large body of research has studied how technologies such as games, virtual or augmented reality, AI, and learning analytics have been used for developing classroom-based assessments. In this section, the authors examine critical developments in technology-driven, classroom-based innovative assessment practices, which is done using a framework organized around (a) assessment of complex constructs, (b) assessment functionality related to evidentiary inferences, and (c) automaticity of assessment tasks. The authors aim to identify the critical roles that emerging technologies play in assessment practices related to these three dimensions. Of particular focus is emerging computational tools, such as educational learning analytics (ELA) and AI, and their role in the form and functionality of classroom-based assessment.The section concludes with how prior work and emerging technologies come together to point toward future directions in classroom-based assessment development, implementation, and research. Of particular importance will be the role of teachers and teacher PD in harnessing the potential of these emerging assessment technologies in classrooms.

Conclusions. The final section of this report contains a set of *conclusions* based on content from the five major sections and includes *implications* for future research and development on classroom-based assessment in STEM education. The report's contents are intended to be useful and usable for multiple audiences, including members of the CADRE community, NSF program directors and project officers, and STEM education practitioners and policymakers. Hopefully, the contents of this report can help to accomplish two things. The first is to stimulate dialogue among members of these communities that will enhance implementation of effective assessment practices in K–12 STEM education classrooms while also influencing the content and execution of in-service and preservice teacher professional learning programs. The second is to chart a course for high-priority areas for the next decade for STEM classroom assessment research, development, and implementation to be funded by NSF, other federal agencies, and private foundations.

Theory and Research Impacting Classroom STEM Assessment

What follows are brief summaries of theory and research related to the six major developments described earlier. No attempt has been made to be comprehensive regarding relevant published literature. Rather, the goal is to provide capsule descriptions of key developments related to each specific assertion made earlier with pointers to selective supporting literature, including the five major sections of this report, where one can obtain a deeper analysis of relevant conceptual and empirical work.

1. Changes in Standards and Expectations for STEM Proficiency

The last two decades have seen substantial and growing interest in changing the landscape of education through ideas labeled as *deeper learning* and *21st century skills*. This global trend is indicative of a long-standing concern in education about the difficult task of equipping individuals with transferable knowledge and skills (e.g., Bellanca, 2014; Pellegrino & Hilton, 2012). Much of the discussion of deeper learning and 21st century skills has been couched in terms of broad, transversal competencies using generic labels such as *problem-solving*, *critical thinking*, *creativity*, *collaboration* etc. Arguments have been made, however, that deeper learning and the development of 21stt century competencies do not happen separately from the disciplinary content to be learned (see, e.g., Pellegrino & Hilton, 2012). Rather, deeper learning is the product of interconnected cognitive, interpersonal, and intrapersonal processes that enable students to grasp disciplinary content and recognize when, how, and why to apply that knowledge for critical thinking and problem-solving (Pellegrino & Hilton, 2012).

Perhaps it is not surprising that concepts related to deeper learning and 21st century skills are reflected in the disciplinary frameworks and standards introduced over the last 15 years for mathematics and science education. For example, both the mathematics and science standards emphasize using and applying knowledge in the context of disciplinary practices, which are the everyday ways of knowing and doing what mathematicians and scientists employ in their respective fields (National Governors Association [NGA] Center for Best Practices & Council of Chief School Officers [CCSSO], 2010; NGSS Lead States, 2013). The argument for incorporating disciplinary practices into instruction is that "learners, much like professionals, are more likely to advance or deepen their understanding when they have opportunities to use and apply knowledge to solve problems, reason with evidence, and/or make sense of phenomena" (Harris et al., 2019. p. 54). The contemporary mathematics and science standards and frameworks have, in turn, served as a source of inspiration for recently developed disciplinary standards and "big ideas" for emergent K-12 STEM subjects, such as computer science.

The STEM standards have profound implications for what constitutes proficiency in each STEM discipline, how that develops over time with instruction, and what constitutes evidence of that proficiency. An ongoing challenge is how to design curricular materials to support development of these important competencies and how to organize classroom instruction, and especially the design and use of assessment, to promote student attainment of the complex disciplinary objectives embodied by contemporary STEM standards (see e.g., Pellegrino et al., 2014). The implications for what to assess and how to do so, whether at a classroom level or a large-scale assessment level, are profound. Multiple sections of this report confront these issues and offer examples of assessment solutions for application in the STEM classroom.

2. Theories, Models, and Data on Disciplinary Knowing and Learning

The past two decades include publication of important syntheses regarding the general nature of learning and knowing as well as the specifics of learning and knowing in the STEM disciplines (e.g., National Research Council [NRC], 2000; 2018). Within that broader literature, it has been argued that sociocultural theory provides the most compelling general theory of knowing and learning because it accounts for how meanings, purpose, values, and motivation are jointly developed as part of learning and helps to understand the ecology of classroom instruction and learning (e.g., NRC, 2018; Shepard et al., 2018). In this perspective, learning occurs as individuals jointly participate in the practices of a community, using the language, tools, and other cultural artifacts of that community. Accordingly, learning is itself situated within communities and emerges from the practices taken up in different settings and communities (Pellegrino & Hilton, 2012). In the case of school, a classroom constitutes such a community, and learning is situated in the practices of that community for the teaching, learning, and assessment of STEM disciplines such as mathematics and science.

The sociocultural perspective has implications for how the STEM disciplines should be taught and learned in school settings. For example, the disciplines are distinct communities that each engage in their own shared practices of knowledge generation, communication, and revision (Shepard et al., 2018). Thus, it is not surprising that a sociocultural perspective is reflected in the most recent disciplinary frameworks and standards for STEM education. For instance, the science framework and standards call for integrated development of science practices, crosscutting concepts, and core ideas of the discipline. Proficiency, from the science education community perspective, involves much more than being able to recall core ideas; it requires application of those ideas to explain natural phenomena and solve problems using the kinds of practices and habits of mind that disciplinary experts typically employ. A similar argument can be made about the mathematics standards with their articulation of both mathematical practices and core disciplinary content and their integration in solving problems and constructing explanations.

Sociocultural theory has multiple implications for assessment in the classroom at a broad level, including the role that assessment plays in the culture of the classroom learning environment, how students come to understand the processes and the practices that constitute it in the classroom, and how it is used by students and teachers to advance their trajectories of learning. Sociocultural theory holds that in addition to uncovering students' disciplinary knowledge, a key role for assessment is to elicit students' relevant interests and experiences and their identities to help inform instructional practices that can help them develop identity related to the disciplinary community (Shepard et al., 2018). This framing of learning has major implications for issues of equity and justice as discussed in **Section 3**.

While sociocultural theory establishes some broad principles about the nature, roles, and uses of assessment in education in general, and classrooms in particular, the specifics of assessment need to be supplemented by discipline-specific models of learning. Currently, learning progressions are the best example of the more detailed disciplinary learning models that Penuel and Shepard (2016a, 2016b) identified as socio-cognitive models of learning. Such models attend to the social nature of learning and to discipline-specific ways that core ideas and practices are developed over time. A variety of definitions and instantiations of the learning progression/trajectory construct exist in the literature, with substantial variability in focus and intent (see e.g., Confrey, 2008; Confrey et al., 2008, 2009; Corcoran et al., 2009; Daro et al., 2011; Duncan & Rivet, 2018; Duncan & Hmelo-Silver, 2009). Two thorough discussions of the learning progression/ trajectory construct and its implications for curriculum, instruction and assessment can be found in the Consortium for Policy Research in Education (CPRE) reports Learning Progressions in Science (Corcoran et al., 2009) and Learning Trajectories in Mathematics (Daro et al., 2011). As described in these CPRE reports (as well as others, see e.g. Rogat et al., 2011), learning progressions and trajectories are empirically grounded hypotheses about how students' understanding and use of core concepts and related disciplinary practices become more sophisticated over time with appropriate instruction. These hypotheses describe the pathways students are likely to follow as they move from emerging understandings to more advanced disciplinary knowledge and practice. They are driven by learning theory and grounded in prior research about how students' learning progresses within domains. The reports emphasize that hypothesized learning progressions and trajectories should be rigorously tested to "ensure their construct validity (Does the hypothesized sequence describe a path most students actually experience given appropriate instruction?) and ultimately to assess their consequential validity (Does instruction based on the learning progression produce better results for most students?)" (Corcoran et al., 2009, p.15). The reliance on empirical evidence distinguishes learning trajectories from traditional topical scope and sequence that describes the content coverage and order in which the content will be taught. Quite differently, a scope and a sequence are typically based on a logical analysis of current disciplinary knowledge, constraints of the school calendar, and intuitions derived from experiences in teaching.

STEM education literature has embraced the concept of learning progressions. For example, the available literature on learning progressions was used in developing the Common Core State Standards in Mathematics (CCSS-M), *A Framework for K–12 Science Education*, and the Next Generation Science Standards (NGSS). A considerable body of empirical research has emerged in both science and mathematics focused on developing and testing such progressions for various aspects of the standards. As such, the work varies considerably in scope, content, and specificity. Many progressions are derived from cross-sectional data spanning multiple grade levels while others are based on longitudinal data from within a single grade level or instructional topic. Consequently, progressions vary in their levels of detail and specificity, or "grain size," at which they describe elements along the progression with varying implications for instruction and assessment. **Sections 1**, **2** and **4** discuss various implications of learning progressions for classroom STEM assessment and provide examples of their use.

Even when learning progressions have a strong empirical basis, and systems of curriculum, instruction, and assessment have been developed and validated, caution should be observed in implementation. It does not follow that all learners in all contexts will progress reliably along a progression as specified. To be appropriately sensitive to the varied ways in which learners may progress, learning progressions may require adaptation in local contexts (Lehrer & Schauble, 2015; Penuel, 2015). When developed as part of a progression, assessment tasks or guiding instructional questions should elicit and illuminate student thinking so that students' ideas are integral to what is acted upon in instructional activity (Shepard et al., 2018). Importantly, such tasks or instructional questions should be designed to help students recognize disciplinary ways of thinking and relate them to their own interests and ways of thinking (Bang & Medin, 2010).

Section 3 builds on these sociocultural framings to explore how STEM assessment can center equity and justice. It defines equity and justice following the National Academies of Sciences, Engineering, and Medicine (NASEM, 2022) report on science and engineering in elementary grades, whose subtitle *The Brilliance of Children and the Strengths of Educators* highlights asset-oriented perspectives on both learners and educators. While projects moving toward equity—seeking to broaden access to STEM learning—are increasingly common, projects moving toward justice—expanding what constitutes STEM and linking up with larger social movements—are still emerging.

3. Coordination and Integration of Curriculum, Instruction, and Assessment

While the focus of this report is assessment in the STEM classroom, assessment is one of three central components of the complex ecological system of classroom teaching and learning that includes curriculum, instruction, and assessment. The three functional components are intertwined, although how they link and work together to influence what happens in classrooms is often less explicit than it should be (Pellegrino et al., 2014). Moreover, the connections among the three are often inconsistent in practice, which can contribute to an overall incoherence in the learning context of the classroom and more broadly in the larger educational system.

Curriculum consists of the knowledge and skills in subject matter areas, often instantiated in curriculum materials, that teachers teach, and students strive to learn. The curriculum generally consists of a scope or breadth of subject area content, a sequence for learning, and a set of resources (e.g., lesson plans and curricular units) that equip teachers and students to accomplish the goals of learning. Content standards typically specify the goals of learning in a subject area, whereas, curriculum sets forth the means to be used to achieve them. *Instruction* refers to the actions taken in classrooms to create conditions for learning to occur. These actions include the methods of teaching and the learning activities used to help students advance in their learning relative to the content and learning goals specified by a curriculum. It encompasses the activities and interactions of teachers and students with one another and often includes interactions with curricular resources. In classrooms, it is through instruction that students engage with the content to be learned, and this engagement can be mediated by curriculum materials. *Assessment* is the means used to evaluate the outcomes of instruction and monitor the progress of students in achieving important learning goals. When directed toward these aims, assessment may include both formal methods, such as large-scale assessments and districtwide or schoolwide exams, or less formal classroom-based procedures, such as quizzes, class projects, and teacher questioning. Whereas curriculum and instruction shape students' opportunity to learn, assessment is the means to measure the intended outcomes. Noteworthy is that assessment can also be used as a tool for improving instruction and advancing students' progress in achieving learning goals. Although less common, this use of assessment shifts the emphasis away from evaluation and grading and toward using assessment to inform teaching and to support ongoing learning. When used in this manner, assessment becomes more classroom focused with potential benefits for day-to-day instruction and for continuous improvement of learning. This format of assessment may include performance tasks as well as projects or products from instructional activities.

Section 1 elaborates on the argument that STEM educational practice is most impactful when there is alignment among curriculum, instruction, and assessment; meaning that the three should work together and toward the same aims, rather than work at cross purposes. Ideally, an assessment should measure what students have an opportunity to learn, and what is taught via instruction should parallel the curriculum and goals of learning. If they are not well synchronized, the balance of the learning environment will be disrupted, and the learning process thrown off. Accordingly, assessment results will be misleading and/or instruction will be suboptimal. Alignment among curriculum, instruction, and assessment can be best achieved if all three are designed from a scientifically credible and shared knowledge base about cognition and learning in subject matter domains (McCallum & Pellegrino, 2018). Without such a central conceptualization of disciplinary learning, the pressure to prepare students for high-stakes tests becomes intensified, and teachers may feel compelled to move back and forth between instruction and external assessment and teach directly to the items on high-stakes tests. This approach, in which external assessment drives classroom instruction, can result in an undesirable narrowing of the curriculum with limited learning outcomes. Such problems can be avoided if, instead, decisions about both instruction and assessment are guided by models of learning that represent the best available scientific understanding of disciplinary knowing and learning in the domains. **Section 1** elaborates on the role of models of learning such as learning progressions in the process of integrating and coordinating curriculum, instruction, and assessment, and provides examples from the STEM research portfolio. **Section 3** builds on this theme of integration and highlights that all three need to be aligned to fully support equity and justice in STEM education.

4. Assessment as a Process of Reasoning from Evidence

Although assessments take several forms and are currently used for many purposes in STEM classrooms, the aim of assessment should be "*to educate and improve* student performance, not merely to *audit* it" (Wiggins, 1998, p.7). The goal in assessing students is to learn what they know and can do and use it productively. But assessing educational outcomes is not as straightforward as measuring height or weight; the attributes to be measured are mental representations and processes that are not outwardly visible. Thus, assessment is designed to elicit students' behavior to produce evidence that can be used to draw reasonable inferences about what students know. Collecting evidence in support of inferences about what students know represents a chain of reasoning based on evidence about student learning that characterizes all assessments: from classroom quizzes to standardized achievement tests and from AI tutors to ad hoc conversations a student has with their teacher as they work through a task or problem.

In the 2001 National Research Council report, *Knowing What Students Know: The Science and Design of Educational Assessment*, the process of reasoning from evidence was portrayed as an *assessment triangle* of three interconnected elements (NRC, 2001). The first element represents a model of student *cognition* and learning in the domain of the assessment. The second element represents a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies. The third element represents an *interpretation* process for making sense of the evidence. These three key elements underlie any assessment and are positioned at the corners of a triangle because each is connected to—and dependent on—the other two. A major tenet of the *Knowing What Students Know* report is that for an assessment to be effective and valid, the three elements must be in sync given the intended interpretive purpose.

The *cognition* corner of the triangle refers to what is known about how students represent knowledge and develop competence in a subject matter domain (e.g., functions in algebra or chemical reactions in physical science). When designing for any particular assessment, a model of learning in the domain is needed to identify the set of knowledge and skills that is important to assess for the context of use, whether that be characterizing the competencies students have acquired at some point in time to make a summative judgment or for making a formative judgment to guide subsequent instruction so as to benefit learning. Importantly, the model should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise.

Every assessment task is based on principled assumptions that the behaviors and/or artifacts the task elicits are evidence of the targeted knowledge and skills. Thus, the tasks to which students are asked to respond to must be carefully designed to provide actionable evidence linked to a model of learning and the associated inferences and decisions. The *observation* corner of the triangle centers on the design of assessment tasks that will elicit these kinds of responses from students. With well-designed tasks, there exists the opportunity to make robust observations, increasing the value of evidence collected and raising the certainty of what we know about a student's knowledge and abilities.

As noted above, every assessment is based on underlying assumptions and models for interpreting observational evidence. The *interpretation* corner of the triangle represents the methods and tools used to reason from necessarily fallible observations. This corner expresses how observations emergent from assessment tasks represent evidence about targeted knowledge and skills. With large-scale assessment, the interpretation is typically done via a statistical model summarizing predicted patterns of student competency. In the context of classroom-based assessment, the interpretation is often less formal, done in a qualitative, implicit fashion by the teacher.

What is important to remember is that each of the three elements of the assessment triangle must both make sense on its own and relate to the other two elements in a coherent and sound way such that it results in an effective assessment and grounded inferences. Underlying the entire process are theories, models, and data on how students learn and what they know as they develop competence. Multiple sections of this report illustrate this *reasoning from evidence* process and provide examples of the ways in which models of learning can be used to guide development and integration of the three elements of the process—cognition, observation, and interpretation—to provide valid classroom assessments for STEM disciplines.

5. Classroom Assessment Practice and Teachers' Assessment Literacy

The literature on educational assessment covers many issues involving both large-scale standardized assessment as well as classroom assessment. Part of that larger literature, generally falling under the heading of *teacher assessment literacy*, is focused on teacher knowledge of assessment and its relationship to classroom practice (e.g., Pastore & Andrade, 2019). Early attempts to define teacher assessment literacy were largely focused on knowledge of measurement theory and practice as applied in large-scale standardized tests. However, the widespread diffusion of information and arguments about the importance of assessment for learning (e.g., Andrade & Heritage, 2017; Black & Wiliam,1998; Wiliam, 2010) has produced a considerable shift in the way assessment literacy has been conceived and the implications for teacher learning and development. The research and development literature on teacher professional learning and classroom assessment has increasingly focused on helping teachers understand and differentiate between the formative and summative functions of assessment and acquiring knowledge and expertise in implementing the elements of good practice for both types of classroom assessment. **Section 4** covers many of the elements of this shift in focus and the complexities it introduces in terms of teacher assessment literacy and programs to help teachers master aspects of effective classroom assessment practice.

In part, the emphasis in teacher assessment literacy on differentiating between the formative and summative functions of classroom assessment is a reaction to a somewhat pervasive but inappropriate dichotomy in the assessment literature conflating assessment purpose and context. Often it has been assumed that classroom assessment is synonymous with the formative use of assessment and that external, large-scale assessment is synonymous with the summative use of assessment. In fact, the summative and formative functions can be identified for most assessment activities regardless of the context in which they occur. Helping teachers understand that classroom assessment is not by definition and execution automatically formative, and that it has historically been practiced as summative is one aspect of developing teacher assessment literacy.

Another aspect in the development of assessment literacy is moving beyond the simple dichotomy between *internal* classroom assessments administered by teachers and *external* tests administered by districts, states, or nations. Ruiz-Primo et al. (2002) argued that these two very different assessment contexts are better understood as two points on a continuum that is defined by their distance from the enactment of specific instructional activities. Ruiz-Primo et al. identified five discrete points on the continuum of assessment distance: *immediate* (e.g., observations or artifacts from the enactment of a specific classroom activity), *close* (e.g., embedded assessments and semiformal quizzes of learning from one or more activities), *proximal* (e.g., formal classroom exams of learning from a specific curriculum), *distal* (e.g., criterion-referenced achievement tests such as required by the U.S. Every Student Succeeds Act), and *remote* (broader outcomes measured over time, including norm-referenced achievement tests and some national [e.g., National Assessment of Educational Progress or NAEP] and international achievement measures [e.g., Programme for International Student Assessment or PISA]). Different assessments need to be understood as different points on this continuum, especially with respect to their capacity to fulfill the formative and summative functions of assessment through their alignment with curriculum and instruction.
The timescales for the five levels defined above can be characterized as *minutes, days, weeks, months,* and *years*. Timescale is important because of the different competencies that various assessments assume and, therefore, their capacity for contributing information to fulfill the timescale-specific formative or summative functions of assessment.

6. Technology and Assessment: Affordances and Opportunities

There has been considerable speculation and writing over the years about the potential of technology to transform assessment for both classroom-based formative and large-scale summative purposes (see e.g., Behrens et al., 2019; Gane at al., 2018, Means et al., 2019; Pellegrino & Quellmalz, 2010; NRC, 2001). For this present report, two considerations are most salient regarding use of technologies for STEM classroom assessment: (a) what can be validly and reliably assessed and (b) using the results of assessment formatively to inform instruction and guide student learning.

Utilization of technology does not change the importance of adhering to contemporary conceptions of student *cognition* in a STEM domain, including the expectation that learners should be able to leverage their disciplinary core knowledge to engage in domain-specific practices while solving problems (NGA, 2010; NRC, 2012). However, the continued expansion of what students are supposed to know and be able to do provides an opportunity for technology to enhance the *observation* and *interpretation* components of the assessment triangle.

With regards to *observation*, technology provides opportunities for the presentation of dynamic and interactive graphic assessment stimuli that have the potential to elicit a broader, more revealing set of responses from students. In addition, technology allows for the integration of student responses over time, space, and modalities to paint a richer picture of student abilities. These emerging technological capabilities also allow for an adaptive environment in which students can be more effectively probed with regards to what they know and can do, thereby, better matching the intended reasoning and response processes that form the basis for desired claims about student proficiency in STEM (e.g., Gorin & Mislevy, 2013).

In addition to capturing data from across a broader set of modalities, data can also be recorded and analyzed in finer-grained detail. This greater detail has the potential of enhancing understanding of the operations and behaviors students engage in when creating products of assessment. Coupled with technology-driven stimuli, such as simulations, this log data can reveal what they are doing and when they are doing it, providing deeper evidence from which to infer proficiency (see e.g., Ercikan & Pellegrino, 2017). However, the volume of data does not automatically translate into deeper understanding. Thus, the ability to increase the quantity of log data has to match the analytic ability to distill it into actionable information to make judgements about students' knowledge, skills, and abilities (see e.g., Bergner & von Davier, 2019).

The *interpretation* corner offers significant opportunities for technology to assist in the reasoning from evidence process. Some of the more obvious forms of assistance are through mechanisms such as automated scoring of multiple-choice responses, parsing through large data sets, and statistical analysis of response data. However, the potential exists for deeper examination of the global and local strategies students utilize while engaging in assessment tasks. Emerging technological capabilities also hold promise for the analysis of less structured data, such as data associated with constructed response questions where students may be expressing their ideas in written and/or graphical form as an argument or explanation.

The processing and analysis of rich, fine-grained multimodal data is likely to be greatly enhanced with emerging AI and associated machine-learning techniques. Research and development efforts have already demonstrated reliable automated scoring of short written constructed responses for various topics and content in science and other subjects (see e.g., Beggrow et al., 2014). Future work is likely to utilize machine learning to allow researchers to analyze more complex response process data of the type described above (Zhai, 2021). Such data may prove to be especially useful in understanding how student thinking and reasoning unfolds as they move through a problem space and thus add to our understanding of current and developing student competence.

Section 5 discusses many of the above noted affordances and capabilities of technology and AI in greater detail and provides examples of advances in selected areas that have been designed to support effective STEM classroom assessment practice. In addition, it addresses one of the most significant issues for teachers regarding implementation of classroom formative assessment. Even when they understand and embrace formative assessment as an important part of their classroom practice, teachers still encounter operational challenges in implementing the process. A major operational obstacle is the burden of managing a complex process, which includes (a) selecting and then administering assessment tasks, (b) capturing student responses, (c) evaluating student responses for strengths and weaknesses, (d) interpreting student outcomes relative to trajectories of learning, and (e) considering the implications for instruction. All of this unfolds under the real time and resource constraints of a classroom. Thoughtfully designed technology-based assessment applications could significantly impact the implementation of formative assessment practices and student learning in the STEM classroom if support can be provided for as many of the operational elements of the formative assessment process as possible. **Section 5** discusses ways in which technology and emerging AI tools can address these pragmatic challenges of implementation and execution.

Coda

The effective use of assessment in the STEM classroom to support student learning is more complex and challenging than one might otherwise assume it to be. It requires the synthesis and integration of multiple areas of theory and research regarding the knowledge to be assessed, teacher knowledge and practice, the complex ecology of the classroom, and resources and tools for implementation and management. The following five sections of this report provide multiple perspectives on what has been learned to date for each of these various topics, including their integration, while simultaneously identifying directions for much needed future research and development. The concluding section provides a summary and recommendations emerging from each section along with an overall set of implications for research, practice, and policy.

References

Andrade, H. L., & Heritage, M. (2017). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*(6), 1008–1026.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science education and Technology*, *23*, 160–182.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science*, *683*(1), 217–232.

Bellanca, J. (2014). *Deeper learning: Beyond 21st century skills*. Solution Tree Press.

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Confrey, J. (2008, July). *A synthesis of the research on rational number reasoning: A learning progressions trajectories approach to synthesis*. Paper presented at the 11th Meeting of the International Congress of Mathematics Instruction, Monterrey, Mexico.

Confrey, J., Maloney, A. P., Nguyen, K. H., Mojica, G., & Myers, M. (2009, July). *Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories.* Paper presented at the 33rd Conference of the International Group for the Psychology of Mathematics Education, Thessaloniki, Greece.

Confrey, J., Maloney, A., Nguyen, K. H., Wilson, P. H., & Mojica, G. F. (2008, April). *Synthesizing research on rational number reasoning.* Paper presented at the National Council of Teachers of Mathematics Research presession, Salt Lake City, Utah.

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

Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. Consortium for Policy Research in Education.

Duncan, R. G., & Hmelo-Silver, C. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal for Research in Science Teaching, 46*(6), 606–609.

Duncan, R. G., & Rivet, A. E. (2018). Learning progressions. *International handbook of the learning sciences*, 422–432.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education*, *53*(2), 176–187.

Gorin, J. S., & Mislevy, R. J. (2013, September). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment*. K–12 Center at ETS, Invitational Research Symposium On Science Assessment. **https://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf?_gl=1*god4om*_gcl_au*NDU0ODk5NjE3LjE2ODU4MTc5NDE**

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67.

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education, 99*(3), 432–437.

Means, B., Cheng, B., & Harris, C. J. (2019). How the science of learning is changing science assessment. In P. Kuhl, S. Lim, S. Guerriero, & D. Van Damme (Eds.), *Developing minds in the digital age: Towards a science of learning for 21st century education* (pp.197–207)*.* OECD Publishing.

National Academies of Sciences, Engineering, and Medicine. 2022. *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators*. National Academies Press. **https://doi.org/10.17226/26215**

National Governors Association Center for Best Practices & Council of Chief School Officers. (2010). *Common core state standards for mathematics*.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school (expanded edition).* National Academies Press.

National Research Council. 2001. *Knowing what students know: The science and design of educational assessment*. The National Academies Press. **https://doi.org/10.17226/10019**

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Research Council. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

National Science Foundation. (2020). NSF 20-600: Discovery research preK-12 (DRK-12) program solicitation. **https://www.nsf.gov/pubs/2020/nsf20600/nsf20600.htm**

NGSS Lead States. (2013). *Next generation science standards: For states, By states*. National Academies Press.

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and teacher education*, *84*, 128–138.

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

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. **https://doi.org/10.1080/00461520.2016.1145550**

Pellegrino, J. W., & Hilton, M. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* National Academies Press.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*(2), 119–134.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the next generation science standards*. National Academies Press.

Penuel, W. R. (2015). Learning progressions as evolving tools in joint enterprises for educational improvement. *Measurement: Interdisciplinary Research & Perspectives, 13*(2), 123–127.

Penuel, W. R., & Shepard, L. A. (2016a). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching,* 5th edition (pp. 787–850). American Educational Research Association.

Penuel, W. R., & Shepard, L. A. (2016b). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton, *Handbook of cognition and assessment: Frameworks, methodologies, and applications*. John Wiley & Sons.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational measurement: Issues and practice*, *37*(1), 21–34.

Wiggins, G. (1998). *Educative assessment*: *Designing assessments to inform and improve student performance.* Jossey-Bass.

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. *Handbook of formative assessment* (pp. 18–40).

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education? *Journal of Science Education and Technology*, *30*(2), 1–11.

Connecting Classroom Assessment with Learning Goals and Instruction through Theories of Learning

Leanne R. Ketterlin-Geller, Southern Methodist University, and
Christopher J. Harris, WestEd

1

Classroom assessments are a powerful driver of students' opportunities to learn. Through the intentional and planful use of classroom assessments, teachers receive data that can help them better design their instruction toward deepening, extending, and refining students' learning in STEM disciplines. However, these aims can only be achieved if there is a clear and explicit connection between the design and outcomes of classroom assessments and the teaching and learning process.

This alignment can be facilitated by focusing on the content, task design, and data generated through classroom assessments. To ensure alignment, the content on which the assessments are based should be relevant to instruction; the design of tasks should represent and elicit ways of knowing that honor and reflect the classroom community and practices; and student response data should be organized and communicated in ways that inform teachers' practices and instructional decision-making. Underlying these assertions is the assumption that data accurately reflect students' knowledge, skills, and cognitive processes, and factors such as the content, task design, and output impact the plausibility of such assertions.

In this report section, we discuss the importance of aligning classroom assessments with learning goals and instructional practices to both shape and evaluate students' learning opportunities. We describe a plausible solution for improving alignment by integrating theories of learning in the design of classroom assessments. We discuss ways in which the specification of theories of learning as learning progressions can improve alignment between classroom assessments and instruction by focusing on the content, task design, and data generated from classroom assessments. We illustrate applications by highlighting research projects funded by the NSF's DRK-12 program.

The Role of Classroom Assessment in STEM Teaching and Learning

Assessment refers broadly to the process by which samples of student behavior are collected to make decisions regarding student learning or performance (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). It is a fundamental practice in STEM classrooms and primarily used as a means for evaluating and monitoring learning, although it can also be a powerful tool for improving instruction and advancing students' progress in achieving learning goals. When assessment is used for the latter, it is considered *formative*. Classroom assessments, when used formatively, can benefit instruction and advance student learning by transforming students' opportunities to learn (c.f. Shepard, 2000; Shepard et al., 2018). For example, in mathematics, teachers can use data gathered while students are modeling solutions to situated addition problems using concrete manipulatives to identify students' conceptual understanding of addition. Teachers' observations can inform their next steps in instruction, such as adjusting the number range used in subsequent problems to reinforce or extend the concept, designing additional practice opportunities, or transitioning to modeling addition using visual representations. In science classrooms, teachers can use evidence from students' performance, such as on a well-designed task that requires students to develop and use a model to explain a phenomenon, to promote productive class discussion and encourage rethinking and revision. When students share and discuss models, teachers have an opportunity to direct student thinking toward what makes for a good explanatory model and invite students to collaborate in reviewing models and iteratively improving them.

As these examples illustrate, formative uses of classroom assessment can play a significant role in shaping students' learning opportunities when the data provide teachers with meaningful and actionable information about students' progress toward reaching the learning goals. By meaningful and actionable, we mean information from assessments that can be used by teachers to inform next steps in instruction in ways that address students' needs. To generate this kind of information, it is important to create alignment between the knowledge, skills, and cognitive processes assessed; the learning goals; and the opportunities provided through instruction.

In STEM education, *knowledge, skills,* and *cognitive processes* underlying the learning goals are associated with the disciplinary knowledge, practices, and ways of knowing or reasoning to be learned. They are often considered together as they reflect the intertwined nature of what is known, how it is known, and how it is acted upon (Pellegrino & Hilton, 2012). In mathematics, the NRC's seminal publication titled *Adding It Up* (2001) articulated this interaction as the *Strands of Mathematical Proficiency* and included conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. Further, the CCSS-M distinguished between content standards and the Standards for Mathematical Practice to recognize the integration of content and processes in the learning and knowing of mathematics (NGA Center for Best Practices & CCSSO, 2010). The eight Standards for Mathematical Practice intersect with the content standards; students should engage with the content and practice standards simultaneously to develop deep understanding of mathematics. As such, the learning goals must include both content and practice standards to represent the breadth and depth of students' knowledge in mathematics.

In science and engineering education, the *Framework for K–12 Science Education* (*Framework*) (NRC, 2012) and the NGSS (NRC, 2013) bring to the fore the essential role of using and applying scientific knowledge to deepen students' proficiency. This knowledge-in-use perspective represents a remarkable vision for science learning where today's students are expected to use and apply knowledge in the context of disciplinary practices—that is, the actual everyday ways of disciplinary reasoning that scientists and engineers use in their respective fields—to make sense of phenomena and design solutions to problems. The *Framework* and the NGSS emphasize that all students must have the opportunity to learn and actively participate in science through using and applying the science and engineering practices integrated with disciplinary core ideas in concert with crosscutting concepts to make sense of phenomena or to solve problems. Disciplinary core ideas are the big ideas associated with a discipline, like ecosystems and biological evolution in life science or matter and its interaction in physical science, and which are essential to explaining phenomena. Crosscutting concepts are ideas such as systems thinking and cause and effect that are important across many science disciplines and provide a unique lens to examine phenomena. Accordingly, the learning goals take the form of *performances* that require the integration of all three dimensions. Such performances are known as *performance expectations,* and they are articulated as a set of standards in the NGSS.

What emerges from this description of the knowledge, skills, and cognitive processes underlying mathematics, science, and engineering learning goals is the emphasis on the application of STEM knowledge through complex cognitive and meta-cognitive processes. The importance is echoed in recent calls for classroom practices to mirror meaningful and authentic experiences of the STEM workforce (Pellegrino & Hilton, 2012). However, the challenge remains for educators to design and enact instructional opportunities and assessment tasks that highlight the rich interconnections between multiple dimensions of knowing.

Instruction refers to the interactions among teachers, students, and curricular materials as well as the pedagogical processes in play through which the learning goals are operationalized. Instruction forms the foundation for students' opportunities to reach STEM learning goals. Broadly defined, instruction includes a variety of interactions, actions, and activities that can be teacher-, student-, or community-facilitated and serve the purpose of accelerating learning. These complement other processes that impact learning, such as learner-governed processes of metacognition and self-regulation (National Academies of Sciences, Engineering, and Medicine [NASEM], 2018). Curricular materials have been widely acknowledged for their central role in supporting instruction in STEM classrooms (c.f., Ball & Cohen 1996; Davis & Krajcik, 2005; Stein et al., 2007). They are used by teachers to organize the activities of instruction and to structure learning experiences to help students accomplish specific aims and achieve learning goals. Assessment is recognized as a central means for evaluating the learning outcomes of curriculum and instruction. Far less utilized, however, is assessment information and insights as part of the instructional process to maximally benefit students. When this is done on an ongoing basis in a formative manner, curriculum and instruction can be better calibrated to meet the learning needs of students.

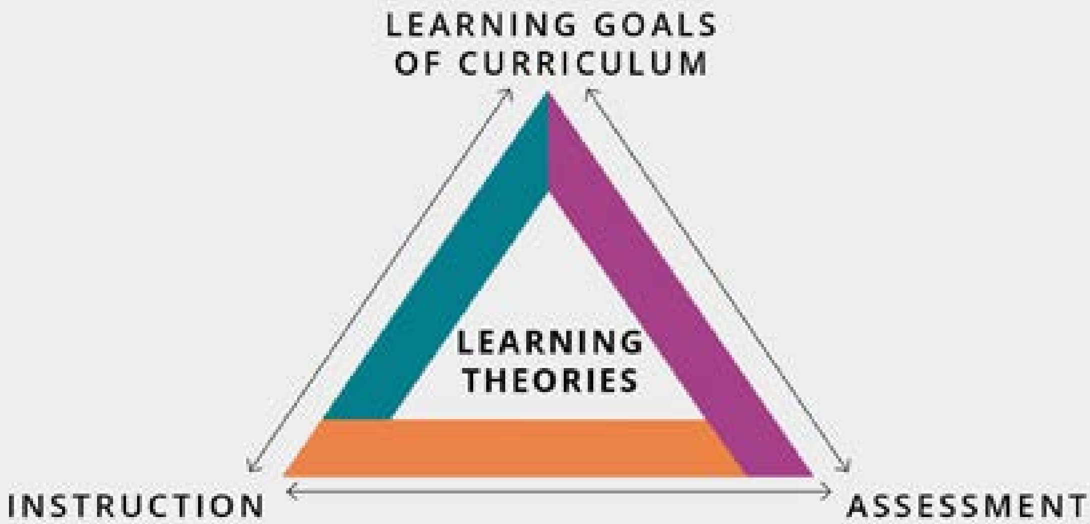
In today's STEM classrooms, assessment has a pivotal role to play in effective teaching and learning. This role is most potent when assessment is integrated into the flow of classroom instruction and used formatively. To realize the use of assessments and assessment data to effectively advance STEM learning will require a tight coupling of assessment with learning goals and instruction.

Critical Importance of Aligning Assessment with Learning Goals and Instruction

The thoughtful aligning of classroom assessments with learning goals and instructional practices has the potential to shape instruction and move students forward in their learning. In this context, the term *alignment* refers to the degree of coherence between the knowledge, skills, and cognitive processes that are elicited by the assessment and those that are identified in the learning goals and/or taught during instruction (AERA, APA, & NCME, 2014). Coherence should be considered across multiple dimensions of the discipline. For example, in mathematics, instruction should provide students with opportunities to learn the content to the same level of rigor as is stated in the curricular expectations. The CCSS-M defines *rigor* as the balance between developing a solid conceptual understanding, building procedural fluency, and applying skills in problem-solving situations. Assessment and instruction should mirror these expectations and provide sufficient opportunities for students to engage with mathematics content through a balance of conceptual, procedural, and problem-solving experiences. In science and engineering education, coherence is emphasized within and across grades through students using all three dimensions of science and engineering practices, disciplinary core ideas, and crosscutting concepts beginning in the earliest elementary grades and upward. Proficiency and expertise develop over time and increase in sophistication as the result of students experiencing coherent and aligned curriculum, instruction, and assessment (NRC, 2012).

A simple but powerful representation of alignment introduced by Pellegrino (2010) and adapted here is depicted in **Figure 1**. This figure illustrates the multi-directional flow of knowledge, skills, and cognitive processes between three central components of alignment. To provide students with meaningful opportunities to achieve the goals of learning typically outlined in standards and instantiated in curriculum, the knowledge, skills, and cognitive processes of instruction and assessment should align with the full depth and breadth of the learning goals. The importance of alignment cannot be understated; classrooms in which there is a high degree of alignment between the learning goals, instruction, and assessment provide students with maximum opportunities to develop competence in STEM domains. Learning theories provide the glue that bonds these three components. When based on robust conceptions of cognition and learning, learning theories help us understand assessment-evoked behavior (guided by cognitive processes) linked to students' knowledge and skills, how these behaviors provide insights as to what they may already know and be able to do, and where they might go from there. Theory also provides guidance as to what forms of instruction might move students toward the established learning goals.

Figure 1. Alignment triangle (adapted from Pellegrino, 2010).[1]



Although the representation in **Figure 1** depicts a straightforward relationship among the three components—for example, an assessment should measure what is taught through instruction and what is taught should be drawn from well-conceived learning goals—it is anything but a simple process to achieve coherence across the three. All too often, there is an imbalance that results in these three components working at cross-purposes. For example, learning goals represent the intended outcomes of instruction, but if the goals are under-specified or the planned instruction is not directed toward the same ends, then alignment can be thrown off, and the assessment will be less effective. In this way, the components are interdependent, and a change or imbalance in one will influence the others. Theories of learning serve as the central core for ensuring that the components are in sync. Fine-grained information about the learning process is needed to guide instructional decisions, such as detailed sequencing of topics, integration of content representations, prior knowledge to activate, and meaningful connections between concepts. Theories of learning can be articulated with such specificity so as to support teachers' instructional decision-making, and thereby present a viable solution for aligning learning goals, instruction, and classroom assessments with the intention of improving students' opportunities to learn.

Theories of Learning and Learning Progressions

The introduction to this report introduced how theories of learning are being used to inform the development of disciplinary learning models that depict how students' learning progresses over time. These models have been instantiated in learning progressions and learning trajectories in STEM education. Although differences exist between learning progressions and learning trajectories (e.g., Delgado & Morton, 2012), their similarities are considerable. Noteworthy is that the term *learning trajectory* is frequently used in mathematics education, and the term *learning progression* is often used in science education, although both are used interchangeably by STEM researchers and practitioners. Confrey (2018), for example, noted in a comprehensive review that the differences are more superficial than substantive. While both are used in STEM education, in this section, we use the term *learning progressions* broadly to also include trajectories.

Learning progressions describe the development of sophistication in students' thinking about discipline-specific topics. The specification of students' thinking should include both the content underlying the topics as well as the cognitive processes through which students interact with the content (practices in mathematics, science, computer science, etc.). By specifying the sequential progression of understanding and performance, learning progressions identify a network of intermediary knowledge, skills, and cognitive processes that lead from foundational to advanced thinking (Bennett, 2015). Because learning progressions outline the processes of learning, it follows that basing instructional design and delivery decisions on learning progressions may facilitate students' understanding in the discipline. Furthermore, aligning classroom assessments with learning progressions may improve the relevance and usefulness of student response data for guiding future learning opportunities. As such, as depicted in **Figure 1**, anchoring instruction and assessment in learning progressions may facilitate alignment.

It is important to note that learning progressions represent hypothesized ways of knowing that reflect a generalized understanding of how students develop greater sophistication in disciplinary topics; they are not fixed or determinant for individual students. Variability will certainly exist within individuals, and the extent of the variability may depend on the grain size by which the learning progressions are specified. For example, if a learning progression describes fine-grained advances in students' knowledge, skills, and cognitive processes, there will likely be more inter-individual variability. Moreover, inter-individual variability may be influenced by prior learning experiences, cultural practices and knowledge, and an infinite number of factors that impact the ways in which individuals construct knowledge. As such, to the extent possible, learning progressions should recognize multiple ways of knowing. **Section 3** discusses important concepts associated with which constructs are prioritized and how students' ways of knowing are recognized; readers are encouraged to integrate these discussions.

Research in this area is emerging. For example, in the DRK-12 CAREER-funded project, *Covariational and Algebraic Reasoning: A New Path to Algebra* (NSF 2142000), mathematics education researchers seek to articulate general learning pathways that support middle school students' understanding of algebraic reasoning by way of their covariational reasoning. They conduct iterative design-based research cycles to first understand how individual students reason about covariation to develop and build an understanding of algebra, and then to synthesize these individual progressions into a generalized pathway that is reflective of multiple ways of developing algebraic knowledge. By recognizing the importance of including multiple ways of knowing and reasoning about covariation, the resulting progression will likely be reflective of diverse approaches to developing understanding. Moreover, this project seeks to identify the contribution of fine-grained understandings of covariational reasoning in the development of algebraic knowledge, which may further support instructional design.

Computer science (CS) education is another STEM arena where work on learning progressions is underway. The K–12 CS community has made a start with the development of learning trajectories for key computational thinking concepts (e.g., debugging, decomposition, algorithmic thinking, and variables) based on CS education research literature (Rich et al., 2017, 2018, 2019, 2022). These trajectories have formed the basis of curriculum design and preliminary work in designing formative assessments in the NSF-funded *Learning Trajectories for Everyday Computing* (LTEC) project (Strickland et al., 2021). In the absence of comprehensive trajectories that cover other aspects of introductory CS learning, Grover and Twarek (2023), used a combination of granular learning goals they outlined and K–12 CS standards (defined by the Computer Science Teachers Association [CSTA]) to aid the design of formative assessments and formative assessment professional learning experiences to support conceptual learning of key introductory programming topics in K–12 CS. Still, much more remains to be done to develop empirically grounded learning progressions to support classroom learning and assessment in K–12 CS.

Learning Progressions as the Basis for Designing Classroom Assessments

Deeply rooted in conceptions of cognition and learning, learning progressions derive from a knowledge base about how learning builds in a domain over time. Below, we describe the ways in which learning progressions can enhance the relevance and usefulness of classroom assessment and illustrate current and ongoing research projects funded by the NSF's DRK-12 program that use learning progressions to guide the design of classroom assessments. We consider three aspects of classroom assessment design where learning progressions can be leveraged: content, task design, and data.

Aligning Content of Classroom Assessments with Learning Progressions

Using learning progressions as the content framework for designing classroom assessments inherently brings greater alignment with the learning process. Both the content and structure of learning progressions facilitates this alignment. Notably, learning progressions specify the knowledge, skills, and cognitive processes that underlie knowing and doing in disciplinary-specific topics. Learning progressions articulate a hypothesized pathway through which these knowledge, skills, and cognitive processes develop as students move from emerging understandings to more advanced disciplinary knowledge and practice. By basing classroom assessments on learning progressions, the content of the assessments may be more sensitive to instruction and better able to elicit how students are advancing toward the learning goals. For example, a classroom assessment that uses learning progressions might include tasks that represent a range of sophistication along a continuum toward the learning goals. Such an assessment could elicit information about students' location in the learning process and the extent to which they are advancing toward the end goals of instruction.

Some learning progressions focus on specifying disciplinary topics in fine detail, while others may take a more coarse-grained approach. For example, the primary aim of the project Exploring K–2 Children Understandings of Visual Representations in Algebraic Reasoning (NSF 2201095) is to articulate learning progressions that specify how children in kindergarten through grade 2 understand visual representations such as tables, graphs, and diagrams of algebraic relations. Because of the narrow scope, this learning progression will likely provide fine-grained information about the development of understanding on these topics. On the other hand, the project titled *Developing and Testing a Learning Progression for Middle School Physical Science Incorporating Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts* (NSF 2201068) outlines learning across multiple middle school grades for physical science concepts. The goal of this project is to articulate the interdependence in learning three core disciplinary ideas (matter, interaction, and energy), two science and engineering practices (constructing explanations and developing and using models), and two crosscutting concepts (cause and effect and systems and system models). Because of the extensive scope of this learning progression, it will likely articulate the learning with less specificity than the former. Regardless of the grain size, most learning progressions identify the entry and exit performances, as well as the intermediary phases through which students' learning progresses. These structural features of learning progressions can be used to inform the design of classroom assessments.

Entry performances mark the lower boundary of the learning progression. These are often conceptualized as the foundational knowledge and skills that students should bring with them as they begin to engage with the focal content of the learning progression (Corcoran et al., 2009). Entry performances are not static or fixed, but instead depend on the specification of the learning progression. In fact, the entry performances of one learning progression may be the exit performances of another, thus linking progressions together in a larger network of learning in a discipline.

Entry performances may include prior knowledge or exposure gained through informal experiences or formal instruction (Duschl et al., 2011), and they may also include emerging understandings, preconceptions, or incomplete knowledge (Alonzo, 2018). There may be considerable variability in the entry performances for individual students. For instance, in mathematics, some concepts may be introduced at home or in out-of-school settings long before they are formally introduced in classroom instruction. Counting by 100s may be such a topic if students have experience with monetary systems that have large denominations (e.g., Mexican peso). Experience with fractions and/or division may also develop earlier than is expected in formal schooling when students are exposed to these topics through cooking, playing games, or sharing with others.

Designing classroom assessments to measure entry performances allows teachers to gather more precise information about students' knowledge, skills, and cognitive processes that are needed to successfully engage in the topic. Teachers can use this information to determine the starting point for their instructional activities. In the likely scenario that students have varying prior knowledge and skills, teachers can form instructional groups to differentiate the initial starting point of instruction. As such, these data can meaningfully support multiple classroom-level decisions that seek to support and improve student learning. Emerging technologies such as AI-driven adaptive technologies seek to support such differentiation by choosing appropriate starting points for students based either on stored knowledge about individual students or engaging students in real-time, ad-hoc formative assessments (See **Section 5**).

Exit performances represent the upper boundary of the learning progression and often describe advanced knowledge and skills that mark the end of the expected learning processes for a given learning progression. Although learning may continue beyond this point, the exit performances often denote the curricular expectations for a specific grade or grade band, canonical disciplinary knowledge, or societal expectations of competence (Corcoran et al., 2009; Duschl et al., 2011).

Designing classroom assessments with the exit performances in mind can help teachers better understand the levels of mastery students will be expected to demonstrate with the learning goals. In turn, students' performance on the assessments can be used to guide teachers' decisions about the sufficiency of their instruction for supporting students' opportunities to meet those goals. Moreover, teachers are better poised to make summative judgments about students' performance and growth.

The entry and exit performances flank the intermediary phases of learning. By their very nature, learning progressions specify the ways in which students develop greater sophistication as they accumulate understanding over time. Intermediary phases of learning reflect the knowledge, skills, and cognitive processes that form a pathway between the entry and exit performances. The intermediary phases are not intended to represent an absolute sequence of understanding that is linear, hierarchical, or conjunctive (Confrey, 2018); instead, they reflect a hypothesized association between knowledge, skills, and cognitive processes that are indicative of greater sophistication in the content. The hypothesized learning pathway will inevitably vary for individual students and may include more or fewer intermediary phases that are uniquely interconnected to reflect the individual's prior experiences, learning, and ways of knowing. And the variability between students will likely increase as the grain size of the learning progression increases. However, the generalized structure may still be useful to guide teaching and learning.

Formative uses of classroom assessments can be designed to measure students' progress in developing understanding by targeting these intermediary phases of learning, and thus, inform instruction that moves students further along the progression. These assessments could be aligned in sequence with a pathway toward greater sophistication to enable teachers to identify where in the learning process students are located. For example, if tasks are ordered to mirror the sequence of the intermediary phases, they should progress from least complex knowledge, skills, and cognitive processes to more complex (Ketterlin-Geller et al., 2019). As students respond to the tasks, they are demonstrating understanding in increasingly more complex content. As Alonzo (2018) noted, structuring assessments to identify students' strengths as opposed to identifying deficits in their knowledge supports an asset-oriented learning environment. If errors emerge or incomplete understandings become apparent, this may indicate areas in which additional instruction can be provided. By providing teachers with information about students' prior knowledge and progress through the learning processes, they have a clear sense of the student's current conceptions from which to build upon for future learning.

In some instances, intermediary phases are articulated to specify the common misconceptions or incomplete knowledge students encounter when learning specific curricular topics. These conceptualizations can be identified through the design of classroom assessments. For example, in a multiple-choice format item, distractors can be analyzed to diagnose specific misconceptions or errors (Ketterlin-Geller et al., 2019). Knowing students' underlying conceptions of the content may help teachers design tasks or learning opportunities that directly address the prior conception to facilitate deeper learning or clarify misunderstandings.

Within the DRK-12 portfolio, several examples exist in which researchers are using these elements of learning progressions to guide the design of classroom assessments. For example, the project titled *Collaborative Research: Middle School Students Graphing from the Ground Up* (NSF 2200777) focuses on middle school students' understanding of graphical representations in STEM fields, focusing specifically on students' understanding of coordinate systems and frames of reference as prerequisites for constructing and interpreting graphs. Through research activities, researchers examine individual students' understanding of coordinate systems and frames of reference and then deductively identify general progressions in students' ways of thinking. These learning progressions then serve as the guide for designing and sequencing tasks that support and build on students' prior understanding to deepen their knowledge of graphical representations.

Aligning Task Design with Learning Progressions

Another aspect of learning progressions that supports the design of meaningful and relevant classroom assessments is the means by which multiple ways of knowing can be specified. As noted, learning progressions can provide fine- or coarse-grained details about how students develop sophistication in their understanding of disciplinary topics. These details can include multiple ways of knowing that honor and reflect diverse approaches to understanding that build on students' prior knowledge to develop greater facility with the content. For example, funds of knowledge (e.g., González et al. 2005) can be integrated with learning progressions to emphasize the ways in which communities and cultures develop and express knowing and understanding about disciplinary topics. By broadening what "counts" as knowing, the processes of teaching and learning may be more meaningful and relevant.

Just as we broaden how knowing is defined, we need to broaden the mechanisms and modes by which students express their knowledge, skills, and cognitive processes. In today's STEM classrooms, teachers need to measure the active and integrated learning that comes from instruction that engages students in using and applying their knowledge for deep reasoning and problem-solving (Harris et al., 2019). Items designed to assess recall of content knowledge will not capture the kinds of robust learning emphasized in contemporary STEM standards. Moreover, the benefits for assessing students' understanding using traditional item formats such as multiple choice (e.g., ease of administration, efficiency of scoring) do not hold for formative uses of classroom-based assessments. Instead, teachers need tasks that are more authentic demonstrations of proficiency and allow for a greater range of expression so they can gain insight into students' performance and use the information to improve teaching and learning.

We turned to the DRK-12 portfolio of projects to identify examples that use learning progressions to design authentic assessment tasks to measure multiple ways of knowing. The project *Investigating the Role of Collaboration on the Development of Student Ideas using a Learning Progression for the Function Concept* (NSF 2101393) explores the role of collaborative problem-solving as classroom assessment tasks. This project uses a learning progression focused on the concept of function for students in grades 9–12 to design tasks that will be completed as part of a three- or four-person team. Mathematical discourse and development of understanding of the concept of function are the intended constructs of the tasks, thereby integrating both standards for mathematical practice and content expectations within the classroom assessment. As this example illustrates, using the learning progression as the content framework for the assessment allows these researchers to use discourse as an authentic method of demonstrating understanding to situate students' current level of understanding within the learning process.

Relatedly, the purpose of the DRK-12-funded project *Collaborative Research: Developing an Online Game to Teach Middle School Students Science Research Practices in the Life Sciences* (NSF 1907437) is to build an online video game for middle school students that focuses on learning research practices within the context of life sciences. In the development of the game-based learning environment, researchers will empirically derive the learning progressions for science research practices. They integrate assessment tasks that are authentic to gameplay to understand how students develop their knowledge of research practices. Students' responses guide the feedback each student receives to support their subsequent learning. This example points to the range of ways students' knowledge can be measured for formative purposes.

Aligning Data from Classroom Assessment with Teachers' Needs

As previously noted, learning progressions are intended to represent how knowledge, skills, and cognitive processes develop as students move from novice to more advanced thinking in disciplinary content and practices. When basing classroom assessments on the content and structure of learning progressions, the data should be reflective of students' current understanding so as to inform subsequent instructional decisions. By providing teachers with information about students' location in the learning process, teachers can determine the instructional sequence that will simultaneously build on students' prior understanding while also providing learning opportunities to progress to the next level of proficiency. Moreover, data from learning progressions-based classroom assessments can also provide fine-grained information about the knowledge, skills, and cognitive processes students have mastered, as well as those that are emerging. These data support teachers' design of learning opportunities, such as lesson content, task design, and practice opportunities.

Data produced from learning progressions-based classroom assessments are substantially different from the data produced from assessment systems that are not based on learning progressions. Data from more typical assessments often report the number of items correct (e.g., raw score, scale score); normative ranking (e.g., percentile score); or performance level (e.g., categorical placement based on cut scores). Although these data may support some instructional decisions, they are less informative for making decisions to guide both the sequencing and design of instruction. Data from learning progressions-based classroom assessments can provide nuanced information about individual student's understanding and performance that teachers need to facilitate learning. Yet, it is important to note that this manner of use of assessment data is different from what is typically done in classrooms. To realize the potential of classroom-based assessment, new models for teacher learning are needed to support teachers in using data for diagnosing and advancing student learning and instructional decision-making.

An example of a DRK-12 project that is focused on providing teachers with meaningful data from learning progressions-based classroom assessments is *Assessing College-Ready Computational Thinking* (NSF 2010265). This project is designed to articulate a learning progression of computational thinking for students in grades 10–12 and the first year of college, and then develop assessment tasks to provide data about students' progress in achieving cognitive and developmental goals. These data are provided to teachers in a timely fashion so they can inform instructional design decisions.

Box 1: Instantiated Example of Classroom Assessments with the DRK-12 Portfolio

The *Measures of Early Mathematics Reasoning Skills (MMaRS)* project (NSF 1721100) illustrates the design decisions involved in creating classroom assessments to improve students' opportunities to learn. The team's design decisions revolved around the goal of providing teachers with data to guide their instruction related to two early mathematics constructs that are predictive of future mathematics and STEM success: numeric relational reasoning (NRR) and spatial reasoning (SR).

The purpose of the classroom assessments developed as part of the *MMaRS* project is to provide kindergarten through grade 2 teachers with information to support two forms of instructional decision-making about students' understanding of NRR and SR. First, teachers use the data to understand students' current conceptualizations of the constructs and then plan future learning opportunities. By understanding students' current conceptualizations, including their location within the learning process and their level of understanding of the content, teachers can design their lessons to extend students' thinking, deepen their understanding, and support their progress. Second, teachers can use data from the *MMaRS* classroom assessments to organize groupings of students based on their lesson objectives. For example, the teacher may organize small groups of students who are at similar locations within the learning process to provide instruction on subsequent concepts. Alternatively, the teacher may create groups of students with varying conceptions of the concepts to collaboratively solve problems. In this way, data from the *MMaRS* classroom assessments are intended to support instructional decisions within the classroom.

To fully realize the intended uses of the *MMaRS* classroom assessments, it was necessary to articulate each construct based on theories of learning. The team used an iterative and empirically based approach to specifying the learning progressions for NRR and SR. In brief, they integrated prior research on children's development of understanding in these constructs with input from mathematics education researchers. Then, they interviewed children ages 5–8 years using open-ended tasks to gather data about the entry and exit performances, as well as top bring greater clarity to knowledge, skills, and cognitive processes that comprise the intermediary phases of learning. Kindergarten through grade 2 teachers were surveyed to tap into their knowledge about how students develop understanding. This information was used to design tasks to elicit student thinking that were subsequently used for think-aloud interviews. Finally, the accumulated evidence was presented to mathematics education researchers to make final revisions to the learning progressions. The goal for using learning progressions to specify the constructs underlying the *MMaRS* classroom assessments was to align the data with the learning process so as to meaningfully inform instruction.

Tasks within the *MMaRS* classroom assessments are designed to illuminate students' reasoning underlying the constructs. Students are first asked questions that align with the knowledge and skills specified in the intermediary phases of the learning progression, and then they are asked about the reasoning processes they used to respond. These reasoning questions allow students to explain their thinking in multiple ways, elicit their prior conceptualizations of the content, and integrate interrelated concepts. The reasoning questions are not evaluated based on "correctness," because there is no one way to correctly reason about mathematics concepts. Instead, they are evaluated on the structure of the argument the student presents to describe their thinking. Because these assessments are intended for students in kindergarten through grade 2, their arguments may rely more heavily on concrete or visual representations of the concepts, such as manipulatives or drawings. Other students may construct their arguments using verbal descriptions, such as creating a story that illustrates the concept. What is evaluated is the extent to which the student generates a mathematically meaningful explanation that justifies their thinking. By allowing students flexibility in the ways in which they express their mathematical understanding, the *MMaRS* classroom assessments intentionally honor and respect the diverse ways in which students learn and express themselves mathematically.

In sum, the content specification and design of the tasks are intended to provide teachers with instructionally relevant information about the depth and breadth of students' knowing for NRR and SR. With this information, teachers can design subsequent learning opportunities to support students' development of understanding in these important early mathematics constructs. Additional research is forthcoming to design score reports that effectively communicate the results to teachers to facilitate their interpretation. Once these reports are co-designed with teachers, additional research is needed to evaluate the situated use of these classroom assessment resources for improving students' opportunities to learn.

Future Directions

There is still much to be learned about aligning classroom assessments with instruction from the perspective of learning progressions. Research over the past 15 years has shown the potential of learning progressions to improve STEM teaching and learning (e.g., Duschl et al., 2011; Shepard, 2018), yet the role they can play in classroom assessments remains an important area for researchers. As learning progressions continue to be mapped out and empirically validated in the STEM disciplines, it will be important to focus research efforts on their use as a framework for developing and using assessments that will inform instructional decision-making. Important questions remain, such as: *In what ways and to what extent do learning progressions-based classroom assessments improve teachers' instructional decisions? How can student response data from these assessments be used to inform teaching actions that will advance student learning? What forms of student feedback will benefit students most for reaching more sophisticated levels of proficiency in a progression?* Moreover, it will be critical to give dedicated research attention to how assessment can and should support meaningful STEM learning for a wide range of students: *How can assessment and instruction work together to support equity and inclusion for all students, especially those who have been historically marginalized from STEM? What are the experiences with learning progressions-based classroom assessments that will engage students in relevant, meaningful, and robust learning? How can equity be incorporated into assessment design so that tasks are relatable to students' varied backgrounds and experiences and accessible and fair for a wide range of students?* This is an important and emerging area for research in classroom-based assessment that should not be ignored. The current NSF DRK-12 portfolio has some assessment projects that are explicitly attending to equity and inclusion; future research should expand on these early promising efforts.

Another direction for research is on new assessment technologies to bring assessment and instruction closer together. There is exciting potential for technology to empower teachers and students to use assessments in new and innovative ways. Recent advances in automated assessment technologies, such as machine learning, can help address the challenges related to efficient scoring and interpretation of student response data. In the not-too-distant future, the benefits of technology-delivered classroom-based assessments that can be used formatively will be realized on a regular basis in classrooms. The assessment research being planned and enacted today should have an eye on the transition of schools to a new generation of responsive digital learning tools where each student has an electronic device to access the Internet and interact with digital learning materials, and teachers have digital access to support their teaching. Importantly, the NSF DRK-12 portfolio should continue to encourage innovative research and development that focuses on assessment technologies that aim to (1) measure the active and integrated learning that comes from robust STEM instruction and (2) support the use of student response data to improve teaching and advance learning.

A final important direction for research is to take on the practical challenges that teachers encounter for classroom-based assessment that have yet to be resolved. For example, a hurdle for many teachers is how to use assessments and assessment data to effectively advance STEM learning. Interpreting data, formulating feedback, and determining appropriate instructional next steps are time-consuming actions for teachers, which create a bottleneck in instruction. The result is that teachers use classroom-based assessment far less often than is optimal for instruction. Moreover, teachers are very familiar with using assessment for grading purposes and far less familiar with using assessment toward improved teaching and learning. When used mostly for grading, teachers miss out on the most powerful benefits of classroom-based assessment. What is needed are assessment projects that build in professional learning for teachers so that they can deepen their knowledge and improve their practice for using assessment in instructionally supportive ways.

References

Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education, 31*(2), 104–112, doi: 10.1080/08957347.2017.1408630

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher*, 25, 8–14.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*, 370–407. doi:10.3102/0091732X14554179

Confrey, J. (2018). *Future of education and skills 2030: Curriculum analysis – A synthesis of research on learning trajectories/progressions in mathematics* (EDU/EDPC[2018]44/ANN3). Organisation for Economic Co-operation and Development.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Consortium for Policy Research in Education.

Davis E. A., & Krajcik J. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3–14.

Delgado, C., & Morton, K. (2012). Learning progressions, learning trajectories, and equity. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann, P. (Eds.), The future of learning: Proceedings of the 10th International Conference of the Learning Sciences (Vol. 1, pp. 204–211). International Society of the Learning Sciences.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education, 47*(2), 123–182.

Grover, S., & Twarek, B. (2023). A formative assessment literacy module for K–12 computer science teachers: Need, design, and teacher feedback. *Proceedings of the 54th ACM Technical Symposium on Computing Science Education (SIGCSE'23).* ACM.

Harris, C. J., Krajcik, J., Pellegrino, J., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice. 38*(2), 53–67.

Ketterlin-Geller, L.R., Shivraj, P., Basaraba, D., & Yovanoff, P. (2019). Using mathematical learning progressions to design diagnostic assessments. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 1–22*.* doi:10.1080/15366367.2018.1479087

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. doi: 10.17226/24783

National Council of Teachers of Mathematics. (2014). *Principles to actions*.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards*.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. National Academies Press.

National Research Council. (2012). *A framework for K–12 science education*. National Academies Press.

National Research Council. (2013). *Next generation science standards: For states, by states*. National Academies Press. **https://doi.org/10.17226/18290**

Pellegrino, J. W. (2010). *The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement.* Center for K–12 Assessment & Performance Management, Educational Testing Service.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Rich, K. M., Binkowski, T. A., Strickland, C., & Franklin, D. (2018, August). Decomposition: A K–8 computational thinking learning trajectory. *Proceedings of the 2018 ACM conference on international computing education research* (pp. 124–132).

Rich, K. M., Franklin, D., Strickland, C., Isaacs, A., & Eatinger, D. (2022). A learning trajectory for variables based in computational thinking literature: Using levels of thinking to develop instruction. *Computer Science Education, 32*(2), 213–234.

Rich, K. M., Strickland, C., Binkowski, T. A., & Franklin, D. (2019, February). A K–8 debugging learning trajectory derived from research literature. *Proceedings of the 50th ACM technical symposium on computer science education* (pp. 745-751).

Rich, K. M., Strickland, C., Binkowski, T. A., Moran, C., & Franklin, D. (2017, August). K–8 learning trajectories derived from research literature: Sequence, repetition, conditionals. *Proceedings of the 2017 ACM conference on international computing education research* (pp. 182–190).

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. NSTA Press.

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165–174.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement:*

Issues and Practice, *37*(1), 21–34.

Stein M. K., Remillard J., & Smith M. (2007). How curriculum influences student learning. In Lester F. K. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 319–369). Information Age Publishing.

Strickland, C., Rich, K. M., Eatinger, D., Lash, T., Isaacs, A., Israel, M., & Franklin, D. (2021, March). Action fractions: The design and pilot of an integrated math+ CS elementary curriculum based on learning trajectories. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (pp. 1149-1155).

Suggested Citation

Ketterlin-Geller, L. R. & Harris, C. J. (2023). Connecting classroom assessment with learning goals and instruction through theories of learning. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives* (pp. 83-94). Community for Advancing Discovery Research in Education (CADRE), Education Development Center.

Assessment for Learning

Arthur J. Baroody, University of Illinois Urbana-Champaign, and
James W. Pellegrino, University of Illinois Chicago

2

Leander, a third grader, had been struggling with mathematics all year. His seatwork, for example, was often incomplete or incorrect. He had made no progress memorizing the multiplication facts. An overall Partnership for Assessment of Readiness for College and Careers (PARCC) minimal mathematics score of 650 indicated he was at Level 1—did not yet meet expectations. An examination of these test responses indicated that he was typically inaccurate, let alone fluent, in producing the product of any two 1-digit numbers from memory. Indeed, the same was true of adding multi-digit numbers. His teacher Ms. Seo recognized that standard drills were not working but was unsure what to recommend to Leander's parents or his next teacher.

The scenario above is played out thousands of times as teachers review their students' year-end achievement test results. In contrast to the more commonly used *summative assessment* (e.g., a unit test given at the end of an instruction unit to assess whether unit content has been mastered or a federally mandated state achievement test to assess year-end attainment of grade-level standards), *formative assessment* is done before and during instruction to gauge learning readiness and to monitor the progress of student learning so as to guide instructional planning (Black et al., 2003; Bloom, 1969; Cizek, 2010; Heritage et al., 2008; Miller & Lavin, 2007; Shepard et al., 2018a; cf. Scriven, 1967). As opposed to *assessment of learning* (summative assessment), formative assessment entails *assessment for learning.* It is characterized by the effort to collect information about the state of a student's knowledge and thinking so instruction can be tailored to further progress.

That said, Bennett (2011) observed there are two perspectives on the definition of formative assessment. One favored by test publishers is that it is embodied as an instrument, such as the Pearson (2005) *Progress Assessment Series* or a curriculum-based measurement (CBM; Fuchs, 2004, 2005; Gersten et al., 2005; Methe et al., 2011; Platas et al., 2016; Shinn, 1995), including probes that identify strengths and weakness with specific skills (Fuchs & Deno, 1991). A broader view favored by many educators and researchers is that "formative assessment is not a test but a process" (Popham, 2008). Based on this view, it is analogous to a verb, not a noun. We favor this latter perspective but allow that, if used thoughtfully to obtain instructionally relevant information, well-designed published tests can be a useful tool in the process of formative assessment.

In an era that emphasizes promoting equity (i.e., the academic achievement of all children), formative assessment is considered an essential ingredient for improving or reforming instruction (Platas et al., 2016; Shepard et al., 2018a). In their seminal review, Black and Wiliam (1998) argued that a lack of formative assessment furthered ineffective teaching and exacerbated the low self-esteem and motivation of low-achieving students. Stiggins (2002) added that efforts to raise standards through standardized summative testing were unlikely to motivate students with low achievement and self-efficacy (cf. Bennett, 2011). See **Section 3** on diversity, equity, and inclusion in assessment for a detailed discussion of these issues.

Central to reform efforts is shifting the focus of instruction to fostering what Hatano (2003) called *adaptive expertise*—meaningful learning that can be applied flexibly and appropriately to new problems (as well as routine tasks)—as opposed to routine expertise—knowledge memorized by rote that may not support appropriate application or transfer (CCSSO, 2010; National Council of Teachers of Mathematics [NCTM], 2006; Pellegrino & Hilton, 2012). Adaptive expertise arises when instruction fosters conceptual, procedural, and factual knowledge in an intertwined or integrated manner that also promotes thinking (e.g., problem-solving and reasoning; NRC, 2001a). In contrast, routine expertise results when instruction focuses on promoting procedural or factual knowledge without understanding, thus limiting the capacity for application and transfer (see also Baroody, 2003; Pellegrino & Hilton, 2012). The related concept of *knowledge in use* is at the core of the NRC (2011) *K–12 Science Education Framework* and the Next Generation Science Standards (NRC, 2013). Like adaptive expertise, knowledge in use focuses on developing science knowledge and reasoning that integrates core disciplinary ideas, crosscutting concepts, and science and engineering practices, thereby supporting knowledge application, including coherently explaining phenomena, and transfer, including solving novel problems (see Harris et al., 2019).

Formative assessment plays a particularly important role in such efforts. In his analysis of why reform efforts aligned with progressive education were not successful, Dewey (1963) concluded that activities, however new and well-intended, were in and of themselves not necessarily educational. Educational activities had to be carefully chosen to ensure external factors meshed with internal factors, such as children's interest and developmental readiness. Consistent with this "principle of interaction," fostering meaningful learning entails identifying what children already know and do not know and building on the former to learn (moderately) new knowledge (Clarke et al., 2014; Claessens & Engel, 2013; Fyfe et al., 2012; Ginsburg, 1977; Piaget, 1964; Vygotsky, 1978).

Formative assessment is an integral aspect of current reform efforts based on a learning progression (LP) (Corcoran et al., 2009; or learning trajectory (LT) (Daro et al., 2011). As described in the Introduction to this report, a learning progression is a theoretically and research-based series of developmental levels of knowledge and understanding. Although the term *learning trajectory* is often used interchangeably with *learning progression*, it will refer here to a relatively fine-grained learning progression used to guide instruction (Lehrer & Schauble, 2015; Lobato & Walters, 2017). Specifically, a learning trajectory has three components: (a) a learning progression, (b) socially important and research-based goals, and (c) theory- and research-based teaching activities to promote attainment of each level (Lobato & Walters, 2017; see also Clements & Sarama, 2008, 2014; Confrey et al., 2012, 2014; Fuson, 2004; Maloney et al., 2014; Sarama & Clements, 2009;). Hereafter, *learning progression* will refer to a stand-alone learning progression; *learning progression-based instruction* will reference instruction guided by a learning progression whether it also includes the other two elements of a learning trajectory (see also Duschl et al., 2011); and *learning trajectory* will denote learning progressions combined with goals and teaching activities.

Formative assessment serves to identify what developmental level a child has already achieved and the next developmentally appropriate level on which instruction should begin (Daro et al., 2011; Shepard et al., 2018a). Moreover, children are regularly assessed during instruction to gauge whether they have mastered a developmental level before instruction proceeds with the next higher level. In sum, Clements et al. (2020, p. 641) noted “the LT approach involves using *formative assessment* (National Mathematics Advisory Panel, 2008; Shepard et al, 2018a, 2018b) to provide instructional activities aligned with empirically validated developmental progressions (Clarke et al., 2001; Duschl, et al., 2011; Fantuzzo et al., 2011; Gravemeijer, 1999; Jordan et al., 2012).” See **Section 4** on teacher’s assessment knowledge and practices for three specific examples or illustrations of learning trajectories.

Recently, learning trajectories have become viewed as a valuable tool for teacher training (Butterfield et al., 2013; Rich et al., 2018). These tools could support learning of subject-content knowledge, pedagogical-content knowledge, content-development knowledge, and formative assessment. For example, Sztajin et al. (2012) argued that “trajectories can support growth in [the] selection of instructional tasks, interactions with students in classroom contexts, and use of students’ responses to further learning” (p. 148).

Issues of Classroom-Based Instruction and Assessment

Issues Addressed by Learning Progression-Based Instruction

Learning progression-based instruction and its integrated formative assessment can be instrumental in addressing five issues of classroom-based instruction or assessment in the domains of science, technology, engineering, mathematics, and computer science (STEM+C). For purposes of exposition and illustration, the learning progression examples that follow come largely from the domain of early mathematics where detailed analyses of learning and development are available. Parallel examples could be presented for other areas of mathematics or in science, albeit without some of the fine-grained nature of early math learning progressions as illustrated below. For example, Lehrer and Schauble (2015) observed that publications such as *Systems for State Science Assessment* (Wilson & Bertenthal, 2006) and *Taking Science to School* (Duschl et al., 2007) have focused attention on the importance of considering long-term changes in students' big ideas–concepts with a central organizing role in STEM+C learning (e.g., matter, evolution, and energy in the domain of science). They further noted that, "because LPs provide a cross-grades perspective on the learning of subject matter knowledge, they have been prominent in recent discussions of assessments, curricula, and standards, including the Next Generation Science Standards" (p. 432; see also Corcoran et al., 2009; Daro et al., 2011).

1. **K–12 classroom instruction and assessment too often lack coordination between levels, such as preschool and the primary grades, or even between grades within a level** (Germeroth, 2022). For example, Leander's third-grade teacher focused on the grade-level Common Core standard of memorizing the products of all basic multiplication facts rather than where Leander was on the long developmental journey to this goal indicated by **Table 1** (Kaufman et al., 2005). Similarly, each of the boy's other primary-grade teachers had focused instructional efforts exclusively on her grade's mathematical goals. Leander's pre-K teachers focused on socialization and language arts.

   With the learning trajectory approach, goals or standards, curricula, and assessment are conceptualized developmentally or longitudinally, and the focus of teaching and testing is on ensuring a child's progress on learning trajectories that span levels (pre-K, primary grades 1–3, intermediate grades 4–6, and so forth) and grades within levels, not that of grade-specific goals. An implication of a learning trajectory approach is that goals or standards, curricula, and assessments at each level (and grades within levels) should be aligned systemwide. As **Figure 2** indicates, within a level (and a grade within a level), assessments should be aligned laterally with the standards or goals and curriculum for that level (or grade). This is analogous to *horizontal coherence* in which curriculum, instruction, and assessment are co-designed and integrated based on the same cognitive theory of learning (NRC, 2001b; Shepard et al., 2018). Moreover, the goals or standards, curricula, and assessments for a level (or grade within a level) should be aligned vertically with the goals or standards, curricula, and assessments, respectively, at levels (or grades) above or below. This is consistent with "*vertical coherence* occurs when stakeholders at different levels of the educational system hold compatible and synergistic visions of learning goals and the means to achieve them" (Shepard et al., 2018a p. 22). Vertical coherence is one of the key principles in the design of the NRC's *Framework for K12 Science Education*. Each science domain is framed in terms of a small set of important disciplinary core ideas that are increasingly elaborated across grades from kindergarten to grade 12. The disciplinary core ideas form the basis for increasingly sophisticated reasoning about phenomena in the natural world. For a detailed discussion of these coherence issues, see **Section 1** of this report on connecting assessment to curriculum and instruction.

   In the learning trajectory approach, for example, Leander's third-grade teacher would have worked in concert with his within-level colleagues (i.e., grade 2 and possibly other primary-level teachers) and pre-K level teachers to ensure prior goals and instruction laid the developmental foundation for the grade 3 goal of memorizing the products of all basic multiplication facts. Readiness assessments each school year would have checked Leander's position on the learning trajectory in **Table 1** and indicated his initial instruction. Periodic assessments would then have checked whether he was ready for the next step in the learning trajectory. If Leander needed to work on grade 2 goals, he would do so with other primary-level children with perhaps the second-grade teacher while his third-grade teacher worked on grade 3 goals with primary-level children ready for such instruction.

2. **A major concern with K–12 instruction is underestimating or overestimating children's existing knowledge.** Without formative assessment, a teacher may be unaware of how diverse their students are developmentally and—like Ms. Seo—teach to grade-level goals ("teach to the middle") with the result that instruction is either too hard or insufficiently challenging for many students. With preschool and kindergarten, particularly, there is often a misalignment between instructional content and pupil development (Balfanz, 1999). Limited knowledge of children's informal mathematical or science development can seriously limit teachers' expectations of young students and lead to a focus on the most basic numeracy content or science conceptual knowledge, which many, or even most, children have already learned (Clarke et al., 2006; Engel et al., 2013; Ginsburg et al., 2008; Kilday et al., 2012; Lee & Ginsburg, 2007a, 2007b, 2009; Li et al., 2015).

   Familiarity with learning trajectories can help educators better define expectations, including those for school readiness. Importantly, formative assessment is central to a learning trajectory approach. Specifically, to ensure developmentally appropriate instruction, such an approach underscores finding children's existing level of knowledge and building on it to promote the next incomplete level of knowledge. Learning trajectories can provide the guidance for carefully connecting formal instruction to preschoolers' informal knowledge (Copple & Bredekamp, 2009; Ginsburg, 1977; Purpura et al., 2013).

3. **K–12 classroom assessment (e.g., unit tests that accompany curricula or achievement tests) too often focus exclusively on skills that can easily be measured based on whether a child's response is correct.** Traditional summative tests too often focus on a narrow range of knowledge and not the key aspects of cognition recommended in contemporary, reform-minded standards, such as well-organized relational or meaningful knowledge and problem-solving and reasoning abilities (NRC, 2001b; Pellegrino & Chudowsky, 2003). As a result, such testing seriously limits their usefulness in planning or adjusting instruction. Pellegrino et al. (2016) provide an example of the limited instructional utility of the end-of-unit assessments found in a frequently implemented K–5 mathematics curriculum. In the classroom mathematics case illustrated earlier, summative testing toward the end of school year did not help Ms. Seo understand Leander's conceptual gaps that contributed to the student's lack of fluency with basic multiplication facts. Specifically, the testing did not reveal whether the boy understood that multiplication is a shortcut for repeated addition and, thus, could build on his existing knowledge of addition. Nor did it reveal whether he had constructed the concepts underlying the number sense critical for the

meaningful memorization of basic addition facts (e.g., recognizing the general rule that adding 1 to a number (*n*) results in the number after *n* in the counting sequence or that the sums of 7 + 5 and 5 + 7 are the same).

As learning progression levels are tied to understanding as well as skills, learning progression-based assessment focuses on the former as well as the latter. This entails attending to process (e.g., examining how a child solved a problem through observation or questioning) as well as product (i.e., the correctness of the child's solution; Ginsburg, 1977, 1997, 2009). See, for example, Point 3 in **Box 2**. On this view, "incorrect" solutions can be informative to both teachers and pupils (Shepard et al., 2018a). For example, systematic errors can reveal misconceptions or how a child's understanding is limited. See, for instance, Point 4 in **Box 2**.

4. **A common blind spot in teaching concerns what to do with a child who is struggling to learn.** Ms. Seo, like many teachers, did not think developmentally when Leander encountered difficulty learning multiplication facts. When this occurs, remedial instruction too often consists of instructional bandages for the immediate problem (e.g., repeating the same instruction or increasing the amount of drill or practice). When learning and assessment are "conceptualized longitudinally," the focus shifts to what is needed to help a child move forward on a developmental trajectory (Shepard et al., 2018a). This includes considering gaps in prerequisite knowledge. By addressing such gaps, this more strategic remedial instruction is far more likely to move a child forward.

   Learning progressions provide a detailed map of what children must already know to make progress, including nonobvious developmental prerequisites. Consider Leander's struggles with memorizing multiplication and addition facts. The learning progression summarized in **Table 1** underscores a possible and nonobvious root cause: not mastering the simplest *n* + 1 facts (the basis for Levels O3 and O4 in **Table 1**, which in turn, stems from not mastering a key counting skill. Specifically, Leander did not automatically know number-after relations (i.e., had not achieved Level C3 in **Table 1**), such as "the number after seven is eight." This prevented his discovery that, for example, the sum of 7 + 1 (adding one to seven) is the number after seven when we count ("seven, eight"). This, in turn, prevented more advanced (automatic) reasoning strategies such as thinking about near doubles such as 7 + 8 as a 7 + 7 + 1 or 14 + 1. Learning progressions, then, guide the process of remedial formative assessment to trace back the origins of a learning difficulty and focus remedial efforts on the root cause of learning difficulty.

   Learning progressions underscore the importance of vertical coherence. Consider an even less obvious implication of the learning progression in **Table 1**—how its initial level *subitizing* (immediately recognizing and labeling a small number of items with an appropriate number word) serves as the key basis for fluency with addition facts, which can then facilitate fluency with subtraction and multiplication (Baroody, 2016; Baroody & Purpura, 2017). Put differently, why should a primary-grade teacher whose goal is fostering fluency with basic facts be interested in what preschool teachers are teaching? Perceptual subitizing of small numbers (1 to about 3) can enable children to see (decompose) larger numbers into smaller ones and learn to recognize the larger via conceptual subitizing (immediate recognition of larger numbers). For example, once a child can perceptually subitize "two," they can decompose a set of four into two groups of two, learn that this is called "four," and—in time—automatize the process so "four" is recognized immediately. This process can then be repeated so that a child can conceptually subitize a set of eight (as four and four) or eventually a set of sixteen (as eight and eight). Basically, playing dice games and dominoes or otherwise practicing perceptual and conceptual subitizing in the preschool and kindergarten years can provide a basis for memorizing the small doubles (e.g., 2 + 2 = 4 and 4 + 4 = 8) and large doubles (e.g., 8 + 8 = 16). Mastery of doubles can serve as a basis for meaningfully memorizing the near doubles (e.g., 8 + 9 = 8 + 8 + 1 = 17 + 1 = 18) and the 2x*n* and *n*x2 multiplication facts. Furthermore, children can similarly recognize part-whole relations such as 4 is 3 and 1 or 5 is 4 and 1, which can foster learning of small add-1 combinations (e.g., 3 + 1 = 4 and 1 + 4 = 5), which in turn, can lead to the discovery of the number-after rule for adding any number and one. Likewise, it can promote other part-whole relations such as number families (1 + 4, 2 + 3, 3 + 2, 4 + 1 all equal 5), which can serve as the basis for the meaningful memorization of subtraction. For example, 5 – 3 = ? can be thought of as the whole five take away the part three, which leaves what other part that added to the part three makes five?

5. **Formative assessment is often unproductive and an additional burden for teachers.** High-quality formative assessment entails three interrelated components: (1) identification of key aspects of content knowledge and their developmental steps based on cognitive theory and research, (2) sound (valid and reliable) tasks and observations for identifying levels of student knowledge, and (3) the knowledge to interpret variations in student knowledge (NRC, 2001b; Pellegrino, 2005, 2018; Pellegrino & Chudowsky, 2003; Penuel, 2015; Shepard et al., 2018a). The first component entails clear goals that involve both concepts and skills and how this knowledge develops in tandem. The second component requires tasks clearly aligned with learning goals and the developmental steps to them. The third component requires professional development that affords the in-depth knowledge of the second. The effective use of formative assessment also requires a fourth component, namely effectively tying the results of such assessments to instruction.

   However, formative assessment too often does not have all four of these components. That is, it is often not clearly linked to well-defined and corresponding goals, assessment tasks, or specific instructional activities. Even if all or some of these elements exist, teachers often have insufficient knowledge to interpret the goals, tasks, and activities accurately. For example, Colorado's efforts to identify how existing kindergarten school readiness measures align with state academic standards raised several concerns (Germeroth, 2022). It was not always clear to content experts what a standard specified, what a particular task measured, or whether or how a task aligned to a specific standard. Even when the goals, test item, and their alignment were clear to content experts, two major problems became obvious. One was that existing measures such as the Teaching Strategies GOLD—Formative Assessment Made Easy

> for Teachers or the North Carolina Kindergarten Entry Assessment (NC KEA) aligned with some standards but not (many) others. Another major problem was that many teachers lacked the professional training to understand the purpose and meaning of a standard or a task or how standards and tasks aligned. These knowledge gaps, in turn, are obstacles to understanding how the state standards and existing assessment tools aligned with existing curricula.

Ideally, the goals or standards, the formative-assessment measures, and the instructional means for achieving each developmental level would be developed interdependently (not in isolation) to ensure alignment—as is done in constructing a learning trajectory. Learning progression-based instruction highlights worthwhile and developmentally appropriate goals, the levels of development that lead to achieving the goals, the pedagogical content knowledge for achieving each step of instruction, and the assessment tasks and/or observations needed to assess progress and determine the next step in instruction (Butterfield et al., 2013; see, e.g., Clements & Sarama, 2013; Murata, 2004; Murata & Fuson, 2006; Wearne & Hiebert, 1988; Wiser et al., 2012). Lehrer and Schauble (2015) noted, for example, "An LP may also draw attention to the value of curriculum materials that inform teachers about the rationale for the design of instruction—an educative curriculum (Davis & Krajcik, 2005)." A learning trajectory approach, then, can provide an efficacious means of professional development (Clarke, 2008; Kutaka et al., 2016; Wilson et al., 2013). As Shepard et al. (2018a) noted, districts need to be "the locus for designing cohort curriculum, instruction, and assessment activity systems because districts have responsibility for curriculum, teacher professional development, and equity; and districts allocate resources for textbooks and assessment" (p. 21).

Practical Barriers to Formative Assessment

There are numerous practical obstacles to classroom assessment, especially if a teacher does not use a learning progression-based curriculum. As Koh et al. (2015) aptly noted, "while the teachers find formative assessment beneficial, they do not have sufficient knowledge of it and face the challenges of heavy workload, time pressure and lack of confidence" (p. 211). See **Section 4** on teacher's assessment knowledge and practices for a detailed discussion of the obstacles teachers face in implementing high-quality assessment, including (a) limited knowledge of developmental psychology, (b) limited strategies for assessing such knowledge, and (c) limited tools that support interpreting assessment results.

Consider, for example, recent and laudable efforts by the Regional Education Laboratory (REL) Appalachia to develop professional development materials to implement Recommendation 1 (teach number and operations using a developmental progression) and Recommendation 3 (use progress monitoring to guide instruction) of the Institute of Education Sciences early numeracy practice guide (Frye et al., 2013). A notable aspect of the REL Appalachia effort is that in order to promote knowledge of developmental psychology, training begins with a developmental progression that has its foundation (initial level) on subitizing and on how subitizing itself (sublevels) unfolds. This effectively identifies and defines the key worthwhile goals of early numeracy instruction and how these goals are, as suggested by developmental and educational research, developmentally related. Another notable aspect of the REL

Appalachia professional development effort is training on progress monitoring and the strategies and tools for implementing formative assessment. The training focuses on helping teachers understand and use a reasonably sound ("Quick Check") task to gauge a child's starting level and sublevel in the developmental progression and their progress through it. The training includes information on possible observations for helping to identify levels and sublevels of student knowledge and to interpret variations in student knowledge. The Quick Check information is subsequently transferred to a Progress Monitoring Chart, which summarizes each child's starting developmental level and sublevel and progress. The chart can then be used to guide instruction, including small-group instruction of children who are at the same or similar developmental level or individualized instruction. Finally, the REL Appalachia effort focuses on tying the result of progress monitoring or formative assessment to information about specific instructional activities and practices.

As challenges faced by the REL Appalachia team to develop a subitizing Quick Check task illustrate, formative assessment is very much a problem-solving effort. Ideally, an initial assessment of subitizing would be brief, accurate, and reasonably reliable.

Solution 1. A possible solution would be a small-group assessment of, say, three children involving one test trial each for sets of 1 to 5 items in which pupils respond to a set shown briefly by holding up fingers. A correct response is recorded on the Subitizing Quick Check sheet as a check (✓) and an incorrect response is recorded as a minus (**–**).

+ **Advantages** include brevity: Only six sessions of a few minutes each are needed to gauge the initial subitizing competence or progress of a class of 18 children. Holding up fingers gives all members of the group a chance to respond independently and quietly. The scoring is straightforward and so easily understood and implemented.

– **Disadvantages** include not directly checking whether children know the appropriate verbal label (a key component of subitizing). Another is that one trial per number may not be reliable or—given the scoring system—valid. For example, for many children, "two" initially means "not 1" or "many." So, scoring a child who responds to 2 items with "two" overestimates the ability to subitize 2 exactly.

Solution 2. Solution 1 disadvantages could be addressed by using two test trials each for sets of 1 to 5 items, recording incorrect responses by indicating a child's actual response (e.g., recording 2 for a response of "two" to a set of 3), and asking children by verbally stating an answer. Note that two correct answers can be interpreted as knowing a number; two incorrect answers, as not. For children who are inconsistent, a tie-breaking trial could be administered. If a number word is overused to label a non-example two times or more, then the child should be considered as not knowing or not subitizing the number.

+ **Advantages** include: Sessions are only somewhat longer of about 5 minutes each, a response consistent with the definition of the construct, and a scoring system that can account for the overapplication of a number word.

– **Disadvantages** include the possibility of invalid results because a child simply imitates a peer's verbal response and a more complicated scoring system.

Solution 3. A possible solution for the disadvantages of Solution 2 would be to *individually* assess children and—to minimize the assessment time and effort—focus on sets of up to 3 items only. However, this solution would have two key disadvantages: (a) individual testing still take perhaps 1.25 hours to test a class of 18 (for a single skill, albeit a critically important one, in one content area) and (b) eliminating trials of 4 and 5 to shorten the testing also eliminates non-examples of 3 for checking the overapplication of the "three."

Clearly, there is much to consider when designing and implementing formative assessment. Any plan has its strengths and weaknesses, and designing an effective plan often requires problem-solving and compromises.

The REL Appalachia example above further underscores that teachers need considerable knowledge to implement formative assessment effectively, including children's development of a content area. This is necessary to understand a goal or standard being assessed, how to score items consistently and accurately, and to determine what instruction would help their students take the next step developmentally. Many teachers lack the professional training to perform the assessment meaningfully and effectively (e.g., accurately and reliably). Bennett (2011) noted, for example, "a teacher who has weak cognitive-domain understanding is less likely to know what questions to ask of students, what to look for in their performance, what inferences to make from that performance about student knowledge, and what actions to take to adjust instruction" (p. 15).

The REL Appalachia example also underscores other practical challenges faced by teachers in effectively using formative assessment. One is assessing and recording each child's initial developmental level and progress through a developmental progression (in each content area). With a small class, the demands on a teacher's time and effort are considerable but perhaps manageable. With larger and larger classes, the demands quickly become enormous and overwhelming for one educator. Even if a teacher observes four children playing a game to, say, collect sufficient information on their subitizing progress, what about the rest of the class?

Plainly, teachers need significantly more support than that usually provided to implement formative assessment. In addition to carefully crafted and implemented professional development, teachers need support and assistance. For example, planning time could enable a teacher to set round-robin learning centers that groups of children rotate through, enabling the teacher to focus teaching and assessment efforts on one of the groups. Teacher aides, parent volunteers, student teachers, or students from upper grades—with some training—could monitor other activities or learning centers and assess progress. Moreover, teachers need doable diagnostic assessment tasks, such as well-designed multiple-choice (or other easily graded) questions that can detect systematic errors or other symptoms of incomplete or inaccurate knowledge.

During the development of the *Test of Early Mathematics Ability* (TEMA, Ginsburg & Baroody, 1983), the Brighton (NY) School District gave its primary teachers extraordinary support by providing each teacher with a substitute teacher for one of the first weeks of school so that teachers could focus on individually administering this test to their class. The feedback from teachers indicated they had not realized how much informal mathematical knowledge their students brought with them to school or their wide range in existing knowledge. This highlights issue 2 raised earlier in the subsection **Issues Addressed by Learning Progression-Based Instruction.** Overall, teachers felt better prepared to differentiate and target instructional efforts to individual needs. Three points are worth noting about this episode: (1) the opportunity afforded the teachers provided valuable professional development; (2) a well-designed summative assessment such as a standardized test can also serve to inform instruction (i.e., serve as formative assessment; Bell & Cowie, 2000; Bennet, 2011); and (3) it underscores Shepard et al.'s (2018a) conclusion that a district needs to take the lead in supporting formative assessment because of its control of resources and responsibility for student learning, teacher professional development, and equity.

A major challenge in assessment for learning, even with good resources and professional learning, is the "management" of the process and the need for tools to make the process more feasible for classroom practice. This is one place where AI-based tools and other forms of automated assessment can be critical. For a discussion regarding issues of their design and classroom or school implementation, see **Section 5. Technology-Based Innovative Assessment.** In brief, policymakers, school administrators, and others interested in improving instruction "need to provide effective and sustained teacher professional development in formative assessment as well as continual teacher support and collaboration" (Koh et al., 2015, p. 211).

What is Known

Learning Progression-Based Instruction

Over the last several decades, educational researchers have used cognitive and educational theories and research to construct many LPs targeting middle and elementary school and even preschool STEM+C (e.g., Catley et al..2005; Clark, 2006; Duncan & Hmelo-Silver, 2009; Duschl et al., 2011; Paik et al., 2017; Shea & Duncan, 2013), including assessment tools (e.g., Alonzo & Steedle, 2009; Gane et al., 2021; Harris et al., 2019; Rich et al., 2018). This is especially true of preschool and elementary mathematics (see, e.g., Clements & Sarama, 2013, 2014; Confrey et al., 2012, 2014; Sarama, & Clements, 2009), where a rich research base on children's mathematical development has been developing for over a half a century (Resnick & Ford, 1981).

According to several reviews, although reformers have embraced learning progressions as an important tool for improving STEM+C instruction, the efficacy and assumptions of learning progression-based instruction have not been directly tested (Frye et al., 2013; Gallacher & Johnson, 2019; Lobato & Walters, 2017; Shavelson & Karplus, 2012; Shea & Duncan, 2013). For example, the *Building Blocks* preschool mathematics learning trajectory promoted numeracy significantly and substantively more than did business-as-usual (BAU) instruction or an intervention organized by mathematical topics (Clements & Sarama, 2008). However, the difference between the *Building Blocks* curriculum and the BAU training may have been due to differences in dosage—participants in the former simply had more numeracy instruction and practice. Although the *Building Blocks* curriculum and topically based intervention were closely matched in terms of content, the superior performance of the former may have been due to other differences, such as different activities or integrated versus discrete content.

A series of 10 experiments were undertaken to rigorously compare learning progression-based instruction to counterfactuals involving either a teach-to-target approach (i.e., instruction that focuses exclusively on the final target level) or the same unordered activities (e.g., control for confounding factors such as dosage; Baroody et al., 2021). Most supported the efficacy and assumptions of a learning progression-based instruction.

Formative Assessment

Several meta-analyses have indicated that inclusion of formative assessment practices, as indicated by medium to very large effect sizes, has a substantial impact on student achievement in various content areas (Black & Wiliam, 1998; Graham et al., 2015; Kingston & Nash, 2011). For instance, in a meta-analysis of 800 studies, Hattie (2009) found that, of 138 factors, formative assessment was the third most influential factor on student achievement. Ozan and Kincal (2018, p. 87) observed that formative assessment's strong impact is derived from the feedback provided both to teachers about their teaching and to students about their learning. Hattie found that feedback was the eighth most overall influential factor on student achievement (see also Lui & Andrade, 2022). As with learning progression-based instruction, training on and the use of formative assessment can be a valuable professional development tool (e.g., Koh et al., 2010).

Like learning progression-based instruction, although formative assessment is frequently recommended and used, relatively little rigorous or high-quality research supports its efficacy. For example, Dunn and Mulvenon (2009) noted: "In reality, a limited body of scientifically based empirical evidence exists to support that formative assessment directly contributes to positive educational outcomes." As with learning progression-based instruction, it is challenging to isolate the impact of formative assessment. For instance, Clarke et al. (2014) randomly assigned 89 at-risk first graders to an experimental treatment that involved formative assessment and a control that entailed business as usual. Although neither group differed on mathematics assessment at pretest, the former group exhibited significantly and—as measured by a large effect size—substantially greater gains. However, although both groups received the same regular classroom instruction, and control participants "were not prohibited from receiving standard district intervention services" (p. 166), a dosage confound might provide a plausible, alternative explanation.

The meta-analyses have been seriously questioned for various reasons (Bennett, 2011; Briggs et al., 2012; Cizek, 2010; Dunn & Mulvenon, 2009; Kingston & Nash, 2011). One is that the methods of meta-analysis were not reported and cannot be critically reviewed. For instance, Black and Wiliam (1998) did not report any quantitative analyses they performed, and the sources of the range of effect sizes cited were not provided. Another problem is "formative assessment" has been operationally defined in various ways, and the research reviewed has been "too disparate to be summarized meaningfully through meta-analysis" (Bennett, 2011, p. 11). Yet another is the varying technical quality of the research. For example, the Meisels et al. (2003) and Rodriguez (2004) studies were observational, and, thus, plausible alternative explanations for treatment effects cannot be discounted.

Box 2: Example from the DRK-12 Portfolio: Development of the Electronic Test of Early Numeracy

Development of the Electronic Test of Early Numeracy (eTEN) Award Number 1621470 and Completing the Development of the Electronic Test of Early Numeracy Award Number 2201939. Principal Investigator: Arthur Baroody. Organization: University of Illinois at Urbana-Champaign. DRL Date: 09/15/2016 – 8/31/2026

The eTEN is an easy-to-use electronic and semi-adaptive test for quickly assessing the numeracy achievement of children ages 3 to 8 years (Baroody & Chen, 2017; Liu et al., 2019). The test can be administered on relatively inexpensive electronic devices such as iPads. The test gauges seven key areas of numeracy: verbal counting, numbering, numerical relations, mental arithmetic, base-ten/place-value, numeral literacy, and multi-digit written arithmetic.

Point 1: The items for each of the seven areas of numeracy assess levels along a learning trajectory. Although the test can serve as a summative measure of numeracy achievement, an examination of item performance provides an indication of where a child might be on learning trajectories tested. Moreover, the eTEN can flexibly be used as a diagnostic test to better pinpoint a child's progress on each of the seven learning trajectories.

Point 2: A game can make assessment more purposeful and engaging (Baroody & Chen, 2017). Consider the "How Many?" task, which entails asking a child to count a collection and—after the child is done counting—asking, "How many [things] did you count?" For many children, such a question seems unnecessary because—to their mind—they have already shown the observer how many [things] there are by counting the collection. Some children, then, find the how-many question confusing and, thus, do not respond. Some children misinterpret the question to mean, "You counted incorrectly, and so count again." For both reasons, the how-many question can underestimate a child's true understanding of the cardinality principle—the understanding that the last number word used in the counting process indicates the total (cardinality) of a collection. The Hidden Stars game is used to assess the cardinality principle. A child is instructed by the teacher: "I'll show a card with some stars on it; count the stars with your finger, then I will hide the stars, and if you tell me how many stars I'm hiding, you win." The game creates a real purpose (in the child's mind) for responding to the how-many question.

Point 3: eTEN is individually administered and can entail strategy identification, so that process (the means of determining an answer) can be gauged. Consider the ability to count a collection in one-to-one fashion—apply the one-to-one principle of assigning each element of a collection one and only one number word. Examining a child's answer ("product")—whether it is correct or not—is by itself NOT sufficient to show that a child understands the principle and can count one-to-one. For example, presented with five stars, Aiden counts five stars "one, two, three, four, five," and Baja counts sub-vocally and announces "six." Although it appears obvious from the children's answers who should and should not be given credit for the competence, an examination of *how* they were determined their answers can indicate otherwise. Aiden counted the first four items correctly, missed the fifth item, and then recounted the second item as "five." Baja labeled each item once and only once while counting sub-vocally counting "one, two, three, four, six." In effect, Aiden answered correctly for the wrong reason, while Baja answered incorrectly but used the correct process. As simply examining product can overestimate or underestimate competence, the eTEN instructions for the one-to-one counting task specify that the child point while counting audibly: "Touch the stars and count out loud."

Point 4: Another safeguard for avoiding false positives is "testing in context." For example, 2-year-old Arianne identified two fingers as "two," which made her parents proud. However, she then identified three, one, four, five, and ten fingers as "two" also, indicating she did not have a conventional understanding of "two" as meaning a pair of items. Many kindergartners and some first graders respond correctly to $n + 1$ expressions such as 4 + 1 and 7 + 1 with "5" and "8," respectively, but then also respond to 4 + 2 and 7 + 3 with "5" and "8," respectively. Similarly, some first and second graders appear to know the "doubles" subtraction facts 12 – 6 = 6 and 14 – 7 = 7, but then also respond to 11 – 6 and 15 – 7 with "6" and "7," respectively. The eTEN (automatically) checks for false positives by comparing a child's responses to (a) two trials involving two items with those involving one, three, and four items to ensure "two" is not overapplied; (b) 7 + 1 with those to 3 + 3 and 4 + 4 to confirm that a "number-after the larger addend" response bias was not used; and (c) 14 – 7 with those to 15 – 9 and 13 – 6 to examine whether a "last-number-heard" response bias was used.

Gaps, Issues, and Future Directions

Learning Progression-Based Instruction

Much research remains to be done on developing, refining, and validating STEM+C learning progressions and learning trajectories (Shea & Duncan, 2013; Lehrer & Schauble, 2015). This is especially true in emerging areas of technology, engineering, and CS education. For example, Gane et al. (2021) observed: "At this time, there is no systematic, evidence-based means to decide *how* elementary students should learn [computational thinking]" (Zhang & Nouri, 2019, p. 143). Indeed, even after several decades of learning progression and learning trajectory development and evaluation in the fields of mathematics and science education, much still needs to be learned about learning trajectory designs in these allied domains. For example, although patterning learning trajectories for early childhood education have been available since Sarama and Clements (2009) published their landmark book on learning trajectories, two recent studies were consistent with Rittle-Johnson et al.'s (2015) view that patterning levels form a "construct map—a probabilistic continuum of knowledge rather than distinct phases of knowledge" (Baroody et al., 2021). More specifically, some developmental levels may be facilitative (not necessary) conditions for higher levels and that some levels may need to be reordered or redefined. Indeed, ladder-like learning trajectories may not be appropriate models of learning in some, or even many, domains (Gallacher & Johnson, 2019; Lesh & Yoon, 2004).

Alonzo and Steedle (2009) noted that validating learning progression-based instruction includes developing items to efficiently assess a student's developmental level on a learning progression. Citing Bell et al. (1985), they further observed that the research used to develop and validate the learning progressions themselves entailed methods, such as the clinical interview, not practical for research at scale or for classroom use. More needs to be done regarding the development and evaluation of practical tools, such as "ordered multiple-choice" items in which each response option corresponds to a developmental level on a learning progression (Anderson et al., 2007; Briggs et al., 2006; Steedle, 2006). More needs to be learned about how questioning practices of small groups or whole classes can serve as an informal but productive means of formative assessment (Ruiz-Primo & Furtak, 2006). Finally, Alonzo and Steedle (2009) implied that issues regarding the role of language in student responses need careful attention. This includes identifying language and cultural barriers to understanding items and designing them to account for language and cultural diversity.

Although research has begun to show that learning progression-based instruction is more efficacious than similar or even the same unordered activities, and assumption of such instruction are valid in a handful of cases, further research is needed to evaluate these conclusions with a broad range of domains within and across STEM+C (Baroody et al., 2021). Research is also needed to evaluate the add-on value of the formative assessment component of learning progression-based instruction on student outcomes and the professional development of teachers. Furthermore, because learning progression-based instruction "need to be supplemented with consideration of obstacles that the student must overcome," much needs to be learned about the obstacles posed by the content itself (e.g., irregularities in the English counting system); instructional materials (e.g., use of number line, a measurement model, to introduce addition and subtraction); and teachers (implying that the equal sign indicates an arithmetic outcome instead of its real meaning as equivalence or "the same as"; Ginsburg, 2009, p. 109).

Formative Assessment

Although extant research "provides some support for the impact of formative assessment on student achievement," clearly there is a "need to conduct research in which more efficient methodologies and designs will lead to more conclusive results and understanding of the impact of formative assessment and evaluation on student achievement" (Dunn & Mulvenson, 2009, p. 9). In particular, "although formative assessment is regarded as a promising way to improve teaching and learning, there is considerable need for research on precisely *how* it influences student learning" and professional development (Rakoczy et al., 2019, p. 154; see also Bennett, 2011; Briggs et al., 2012). Similarly, much is yet to be learned about the internal mechanisms of feedback processing, especially by students (Lui & Andrade, 2022).

Many questions remain about the development of curriculum-based measurement (CBM) or other ready-made tests used to identify critical competencies or otherwise fulfill the mission of formative assessment (Mazzocco, 2005). A key concern needing investigation is to what extent do such tests gauge adaptive expertise (as opposed to routine expertise)? To what extent are such measures coherent and gauge the role of prior knowledge, including informal knowledge, in constructing meaningful knowledge (Confrey et al., 2009; Duschl, 2019; Platas et al., 2016; Van den Heuvel-Panhuizen, 1996). For example, Methe et al. (2011) noted that although children's informal everyday knowledge is generally acknowledged as an important basis for the meaningful learning of formal and largely written knowledge, only about a third of numeracy CBM measures assessed this foundational knowledge. Moreover, what informal and formal items should be included and what balance of these items provides the clearest picture of age-appropriate development for an age group? Does a test gauge how common aspects of children's incomplete (informal) knowledge interfere with formal learning (e.g., counting 3 inches as 4 because the 0 hashmark on a ruler is counted as "one"; i.e., informally interpreting the measurement of the continuous quantity of length as counting a discrete quantity or collection of things)?

Do ready-made tests used for formative purposes gauge key aspects of adaptive expertise—conceptual understanding (well-organized, relational knowledge including that underlying the comprehension of key skills); appropriate application of knowledge (e.g., flexibly choosing the most efficient strategy for a task); and transfer (adjusting known knowledge to solve a novel problem; (NRC, 2001b Pellegrino & Chudowsky, 2003)? For example, could a primary-level teacher infer from a CBM (e.g., Gersten et al., 2005) whether a student could flexibly switch between using a counting-down strategy to solve 9 – 3 in three steps (8 is 1 less, 7 is 2 less, 6 is 3 less) as opposed to counting up to do so in six steps (4 is 1 more, 5 is 2 more, 6 is 3 more, 7 is 4 more, 8 is 5 more, 9 is 6 more) or counting up to solve 9 – 6 in three steps (7 is 1 more, 8 is 2 more, 9 is 3 more) as opposed to counting down in six steps (8 is 1 less, 7 is 2 less, 6 is 3 less, 5 is 4 less, 4 is 5 less, 3 is 6 less)? Could the measure serve to infer whether a student understood that the basic arithmetic facts can be represented as an organized body of related information, including the understanding that addition knowledge could be used to solve an unknown subtraction combination (e.g., 15 – 8 can be thought of as the whole 15 take away the part 8 leaves what part that added to the part 8 makes the whole 15)?

Questions remain about the nature and efficacy of professional development needed to implement formative assessment effectively. A long-standing question is to what extent do teachers need to understand content knowledge to teach and assess it effectively? For example, Ginsburg (2009) made the following observation that is applicable across STEM+C domains:

Knowledge of mathematics—even 'profound understanding of fundamental mathematics' [Ma, 1999]—is essential for teaching and for assessment as well. The point is obvious: how can you teach something you do not understand? But what is not so obvious is that mathematical understanding is required to understand children's thinking. As Dewey [1976] put it many years ago: 'Really to interpret the child's present crude impulses in counting, measuring, and arranging things in rhythmic series, involves mathematical scholarship—a knowledge of the mathematical formulae and relations which have, in the history of the race, grown out of such crude beginnings' (p. 282).

Related questions that need to be addressed are as follows: To what extent do teachers need to understand children's content development, goals and standards, and related instructional activities or strategies to use formative assessment effectively? What degree of professional development is needed to achieve this knowledge and use formative assessment in a faithful, valid, and reliable manner? Ginsburg, for example, argues that clinical interviews as well as observations and tests can be powerful tools for teachers in gathering the data needed for formative assessment. How much training is needed to interview children flexibly and effectively (e.g., devise a task or problem on the spot to test a hypothesis about a child)—let alone use a test or observations to infer children's skill, understanding, and thinking?

Assessment Task Development and Validation

One of the major challenges to implementing effective assessment practices in STEM classrooms is the paucity of rigorously validated assessment tasks that can be used for formative purposes to inform instruction. Validation of instructionally supportive assessment tasks requires rigorous design coupled with multiple sources of evidence supporting assumptions about what the designed tasks are assessing and their capacity to reveal important student differences that are conceptually and instructionally meaningful. As described in the framework developed by Pellegrino et al., (2016), validation of instructionally supportive assessments requires multiple forms of evidence supporting the cognitive, instructional, and inferential validity of the tasks (NSF 0732090, 0918552, 0920242). Such evidence can come from domain expert evaluations of the knowledge required to respond (cognitive validity), empirical data showing systematic variation in student performance (inferential validity), usability of tasks in the classroom by teachers, and interpretability of student performance relative to the model of learning (instructional validity).

The assessment examples for early mathematics provided in this section are the product of a rigorous "reasoning from evidence" process, such as that describe in the introduction to this report. It uses model learning (a learning progression) to specify the characteristics of tasks where the observed student responses will produce evidence that can be interpreted in terms of student knowledge in reference to the LP—what the student knows and can do—with possible implications for instruction. The *Electronic Test of Early Numeracy*, described earlier in **Box 2: Example from the DRK-12 Portfolio,** is the product of such a design and validation process, yielding a tool that can be used by teachers as part of their instructional practice. Examples of such a rigorous design and validation process are available for other areas of the mathematics curriculum and include the NSF-supported work of Confrey and her colleagues for learning trajectories for middle school mathematics (see e.g., Confrey et al., 2021; NSF 1621254).

In science there are far fewer examples of assessments that have been developed and validated for formative or summative use in K–8 classrooms. In part, this arises from the major shift in science proficiency specified in the 2012 *Framework for K–12 Science Education* and the 2013 Next Generation Science Standards. They demand multidimensional integration of conceptual knowledge (disciplinary core ideas and crosscutting concepts) with science and engineering practices and require assessment tasks far different in the performances required than those previously used for science assessment. The latter typically treated content and inquiry as separate strands of science learning for the purposes of assessment and not their integration. The challenges of designing this new generation of assessments were noted in a 2014 NRC report on developing assessments for the NGSS (Pellegrino et al., 2014). That report also argued that particular emphasis should be given to developing assessments for classroom use given the needs of teachers to understand what was expected of students and the implications for their instruction.

NSF has supported a multi-institution collaborative team to develop examples of such assessments for grades 6–8 (NSF 1316903, 1903103, 1316908, & 1316874). A transformative approach was developed for designing classroom-based assessment tasks that could provide teachers with meaningful and actionable information about students' progress toward achieving the competencies associated with NGSS Performance Expectations in the physical and life sciences (Harris et al., 2019). The design model follows the logic of evidence-centered design (Mislevy & Haertel, 2006) and supports systematic development of families of tasks that can be deployed in the classroom during different phases of instruction for a given science topic area. Central to the design approach are sets of learning performances that establish targets to assess student progress toward mastery of the full set of knowledge and competencies required by the performance expectations (Harris et al., 2019). The approach was used to iteratively develop tasks and rubrics aligned with sets of physical and life sciences performance expectations for the middle school grade band. Validation of the assessment tasks was guided by the framework of Pellegrino et al. (2016) for instructionally supportive assessments. Descriptions of the types of validity evidence obtained and outcomes can be found in Harris et al. (2019).

The collaborative team also created an online task portal through which the technology-based tasks can be delivered. A currently supported collaborative team (NSF 1813737 & 1813938) is applying the same design model in collaboration with upper elementary grade teachers to develop classroom assessment tasks and instructional support resources in life, physical, and earth and space sciences for students in grades 3–5 (Zaidi et al., 2022). Information about the combined NSF-supported NGSS design projects can be found at **http://nextgenscienceassessment.org**. All the currently available tasks for grades 3–5 and 6–8 can be accessed at **https://ngss-assessment.portal.concord.org**. To date over 10,000 teachers have established accounts on the portal, and the tasks have been used by tens of thousands of their students. Nevertheless, this is just a beginning, and much more assessment development and validation work are needed for classroom assessments, as well as teacher professional learning to maximize use of these types of resources for the classroom formative assessment process.

Concluding Remarks

Improving instruction requires teachers to focus on fostering adaptive expertise—the integration of procedural *and* conceptual knowledge (CCSSO, 2010; National Mathematics Advisory Panel, 2008)—or knowledge in use (NRC, 2011; NGSS Lead States, 2013). The use of learning trajectories and formative assessment should inform and improve the efficiency of promoting such knowledge. As teachers typically teach as they were taught, and how they were taught too often focused on procedural knowledge in mathematics and content knowledge in science exclusively and summative assessments of this knowledge, reforming instruction and effectively using learning trajectories and formative assessment depends on professional development (Copur-Gencturk, 2015; Ma, 2010). Learning trajectories, formative assessment, and professional development, however, are only as good as the sound research on which they are based, and there is still much to learn about all three and their integration.

References

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421.

Anderson, C. W., Alonzo, A. C., Smith, C., & Wilson, M. (2007, August). *NAEP pilot learning progression framework. Report to the National Assessment Governing Board*.

Balfanz, R. (1999). Why do we teach children so little mathematics? Some historical considerations. In J. V. Copley (Ed.), *Mathematics in the early years*, (pp. 3–10). National Council of Teachers of Mathematics.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–33). Erlbaum.

Baroody, A. J. (2016). Using number and arithmetic instruction as a basis for fostering mathematical reasoning. In M. T. Battista (Ed.), *Reasoning and sense making in the mathematics classroom: Pre-K—Grade 2* (pp. 27–69)*.* National Council of Teachers of Mathematics.

Baroody, A. J. (2017, January). What role should concrete experiences play in early childhood mathematics instruction. In J. Sarama, D. Clements, C. Germeroth, & C. Day-Hess (Eds.), *Interventions in early childhood mathematics education, Volume 53 of Advances in child development and behavior*. Academic Press.

Baroody, A. J., & Chen, Z. (2017, October). *Testing the mathematical achievement of young children using iPad games*. Invited presentation to the Early Childhood Education in the Era of Science and Innovation Conference, Shanghai, China.

Baroody, A. J., Clements, D. H., & Sarama, J. (2021). Lessons learned from 10 experiments that tested the efficacy and assumptions of hypothetical learning trajectories. *Education Science*, *12*. **https://doi.org/10.3390/educsci12030195**

Baroody, A. J., & Purpura, D. J. (2017). Early number and operations: Whole numbers. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 308–354). National Council of Teachers of Mathematics.

Bell, B., & Cowie, B. (2000). The characteristics of formative assessment in science education. *Science Education*, *85*, 536–553.

Bell, B., Osborne, R., & Tasker, R. (1985). Appendix A: Finding out what children think. In R. Osborne & P. Freyberg (Eds.), *Learning in science: The implications of children's science* (pp. 151–165). Heinemann.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. doi:10.1080/0969595980050102

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Open University Press.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. *Teachers College Record, 70*(10), 26–50. **https://doi.org/10.1177/016146816907001003**

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*(4), 13–17.

Brown, C. P, McMullen, M. B., & File, N. (Eds.). (2019). *The Wiley Handbook of Early Childhood Care and Education.* John Wiley & Sons.

Butterfield, B., Forrester, P., Mccallum, F., & Chinnappan, M. (2013). *Use of learning trajectories to examine pre-service teachers' mathematics knowledge for teaching area and perimeter*. **https://files.eric.ed.gov/fulltext/ED572797.pdf**

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution.* National Academy of Sciences Committee on Test Design for K–12 Science Achievement. **https://www.researchgate.net/publication/253384971_Tracing_a_Prospective_Learning_Progression_for_Developing_Understanding_of_Evolution**

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek, *Handbook of formative assessment* (pp. 3–17). Routledge.

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record, 115*(6), 1–29.

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of conceptual restructuring. *Cognition and Instruction*, *24*(4), 467–563.

Clarke, B. A. (2008). A framework of growth points as a powerful teacher development tool. In D. Tirosh & T. Wood (Eds.), *Tools and processes in mathematics teacher education* (pp. 235–256). Sense.

Clarke, B., Clarke, D., & Cheeseman, J. (2006). The mathematical knowledge and understanding young children bring to school. *Mathematics Education Research Journal, 18*, 78–102. doi:10.1007/BF03217430

Clarke, B., Doabler, C. T., Cary, M. S., Kosty, D., Baker, S., Fein, H., & Smolkowski, K. (2014). Preliminary evaluation of a tier 2 mathematics intervention for first-grade students: Using a theory of change to guide formative evaluation activities, *School Psychology Review*, *43*(2), 160–177. **https://doi.org/10.1080/02796015.2014.12087442**

Clarke, D. M., Cheeseman, J., Clarke, B., Gervasoni, A., Gronn, D., Horne, M., McDonough, A., Montgomery, P., Rowley, G., & Sullivan. (2001). Understanding, assessing and developing young children's mathematical thinking: Research as a powerful tool for professional growth. In J. Bobis, B. Perry, & M. Mitchelmore (Eds.), *Numeracy and beyond (Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 9–26). MERGA.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45,* 443–494. doi:10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2013). *Building blocks* (Vol. 1 and 2). McGraw-Hill.

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach* (2nd ed.). Routledge.

Clements, D. H., Sarama, J., Baroody, A. J., & Joswick, C. (2020). Efficacy of a learning trajectory approach compared to a teach-to-target approach for addition and subtraction. *ZDM Mathematics Education, 52*(4) 637–649. **https://doi.org/10.1007/s11858-019-01122-z**

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

Confrey, J., Maloney, A., Nguyen, K., Mojica, G., & Myers, M. (2009). Equipartitioning/splitting as a foundation of rational number reasoning. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Eds.) *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol). 1. Psychology of Mathematics Education.

Confrey, J., Nguyen, K. H., Lee, K., Panorkou, N., Corley, A. K., & Maloney, A. P. (2012). *Turn-on common core math*: *Learning trajectories for the common core state standards for mathematics.* **https://nctm.confex.com/nctm/2013MD/webprogram/Handout/Session19561/Maloney%20NCTM%20RegionalBalt%20TOCC%20101713FINAL.pdf**

Confrey, J., Shah, M., & Toutkoushian, E. (2021, August). *Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone.* In *Frontiers in Education*, 6. **https://doi.org/10.3389/feduc.2021.654353**

Copple, C., & Bredekamp, S. (Eds.). (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8* (3rd ed.). National Association for the Education of Young Children.

Copur-Gencturk, Y. (2015). The effects of changes in mathematical knowledge on teaching: A longitudinal study of teachers' knowledge and instruction. *Journal for Research in Mathematics Education*, *46*(3), 280–330.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Consortium for Policy Research in Education. **https://files.eric.ed.gov/fulltext/ED506730.pdf**

Council of Chief State School Officers. (2010). Common core state standards: Preparing America's students for college and career. **http://www.corestandards.org**

Daro, P., Mosher, F. A., Corcoran, T., & Barrett, J. (2011). *Learning trajectories in mathematics: A foundation for standard, curriculum, assessment, and instruction*. Consortium for Policy Research in Education.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3–14.

Dewey, J. (1963). *Experience and education*. Collier.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*(6), 606–609.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7). **https://www.researchgate.net/publication/237409416**

Duschl, R. A. (2019). Learning progressions: Framing and designing coherent sequences for STEM education. *Disciplinary and Interdisciplinary Science Education Research*, *1*(4), **https://doi.org/10.1186/s43031-019-0005-x**

Duschl, R. A., Maeng, S, & Senzen, A. (2011). Learning progressions and teaching sequences: A review and analysis, *Studies in Science Education, 47*(2), 123–182.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8.* National Academies Press.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157–178. **https://doi.org/10.3102/0162373712461850**

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An integrated curriculum to improve mathematics, language, and literacy for Head Start Children. *American Educational Research Journal, 48,* 763–793.

Frye, D., Baroody, A. J., Burchinal, M. R., Carver, S., Jordan, N. C., & McDowell, J. (2013). Teaching math to young children: A practice guide. National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. **http://whatworks.ed.gov**

Fuchs, L. S. (2004). The past, present, and future of curriculum based measurement research. *School Psychology Review, 33*, 188–192.

Fuchs, L. S. (2005). Prevention research in mathematics: Improving outcomes, building identification models, and understanding disability. *Journal of Learning Disabilities, 38*, 350–352.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488–500.

Fuson, K. C. (2004). Pre-K to grade 2 goals and standards: Achieving 21st century mastery for all. In D. H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 105–148). Erlbaum.

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, *104*, 1094–1108.

Gane, B. D., Israel, M., Elagha, N., Yan, W., Luo, F., & Pellegrino, J. W. (2021). Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades. *Computer Science Education*, *31*(2), 141–168. **https://doi.org/10.1080/08993408.2021.1874221**

Gallacher, T., & Johnson, M. (2019). "Learning progressions": A historical and theoretical discussion. *Research Matters*, *28*, 10–16. **https://files.eric.ed.gov/fulltext/EJ1306758.pdf**

Germeroth, C. (2022). *Final Colorado PK-K assessment state agency advisory meeting*. Marzano Research.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*, 293–304.

Ginsburg, H. P. (1977). *Children's arithmetic: The learning process*. D. Van Nostrand.

Ginsburg, H. P. (1997). *Entering the child's mind: The clinical interview in psychological research and practice.* Cambridge University Press.

Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human* Development, *52*(2), 109–128. **https://doi.org/10.1159/000202729**

Ginsburg, H. P., & Baroody, A. J. (1983). *The test of early mathematics ability* (TEMA). Pro-Ed.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report, 22,* 3–23.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *Elementary School Journal*, *115*(4), 523–547. doi:10.1086/681947

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

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019), Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67. **https://doi.org/10.1111/emip.12253**

Hatano, G. (2003). Forward. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. xi–xiii). Erlbaum.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 metaanalyses relating to achievement*. Routledge.

Heritage, M., Kim, J., & Vendlinski, T. (2008). *From evidence to action: A seamless process in formative assessment?* Paper presented at the American Educational Research Association Annual Meeting, New York.

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology, 104*(3), 647–660. **https://doi.org/10.1037/a0029018**

Kaufmann, L., Delazer, M., Pohl, R., Semenza, C., & Dowker, A. (2005). Effects of a specific numeracy educational program in kindergarten children: A pilot study. *Educational Research and Evaluation*, *11*(5), 405–431.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., Whittaker, J.V. (2012). Accuracy of teachers' judgments of preschoolers' math skills. *Journal Psychoeducation. Assessment*, *30*(2), 48–158. **https://doi.org/10.1177/0734282911412722**

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. **https://doi.org/10.1111/j.1745-3992.2011.00220.x**

Koh, M. H., Hill, J. R., & Barbour, M. K. (2010). Strategies for instructors on how to improve online groupwork. *Journal of Educational Computing Research, 43*(2), 183–205.

Koh, K., Lim, L., & Habib, M. (2015). Building teachers' capacity formative assessment: The Singapore example. *The New Educational Review*, *40*(2), 211–222. **https://doi.org/10.15804/tner.2015.40.2.18**

Kutaka, T. S., Ren, L., Smith, W. M., Beattie, H. L., Edwards, C. P., Green, J. L., Chernyavskiy, P., Stroup, W. W., Heaton, R. M., & Lewis, W. J. (2016). Examining change in K–3 teachers' mathematical knowledge, attitudes, and beliefs: The case of primarily math. *Journal of Mathematical Behavior*, *21*, 147–177. **https://doi.org/10.1007/s10857-016-9355-x.**

Lee, J. S., & Ginsburg, H. P. (2007a). Preschool teachers' beliefs about appropriate early literacy and mathematics education for low- and middle-SES children. *Early Education & Development, 18*(1), 111–143.

Lee, J. S., & Ginsburg, H. P. (2007b). What is appropriate mathematics education for four-year- olds?: Pre-kindergarten teachers' beliefs. *Journal of Early Childhood Research*, *5*(1), 2– 31. **https://doi.org/10.1177/1476718X07072149**

Lee, J. S., & Ginsburg, H. P. (2009). Early childhood teachers' misconceptions about mathematics education for young children in the United States. *Australasian Journal of Early Childhood*, *34*(4), 37–45. **https://doi.org/10.1177/183693910903400406**

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education*, *99*(3), 432–437.

Lesh, R., & Yoon, C. (2004). Evolving communities of mind—Where development involves several interacting and simultaneously developing strands. *Mathematical Thinking & Learning, 6*, 205–226.

Li, X., Chi, L., DeBey, M., & Baroody, A. J. (2015). A study of early childhood mathematics teaching in the U.S. and China. *Early Education and Development, 26*, 37–41. **https://doi.org/10.1080/10409289.2015.994464**

Lui, A. M., & Andrade, H. L. (2022, March). The next black box of formative assessment: A model of the internal mechanisms of feedback processing. *Frontiers in Education*, *7*.

Liu, Y., McCarthy, E., Baroody, A. J., & Eiland, M. D. (2019, November). Analysis of the e-TEN calibration data. In S. Otten, A. G. Candela, Z. de Araujo, C. Haines, & C. Munter (Eds.), *Proceedings of the forty-first annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education* (pp. 1862–1866). University of Missouri.

Lobato, J., & Walters, C. D. (2017). A taxonomy of approaches to learning trajectories and progressions. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 74–101). National Council of Teachers of Mathematics.

Ma, L. (2010). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Routledge.

Maloney, A. P., Confrey, J., & Nguyen, K. H. (Eds.). (2014). *Learning over time: Learning trajectories in mathematics education*. Information Age Publishing.

Mazzocco, M. M. (2005). Challenges in identifying target skills for math disability screening and intervention. *Journal of Learning Disabilities, 38*, 318–323.

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Bickel, D.D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives*, 11(9). **https://www.researchgate.net/publication/256011325_Creating_a_system_of_accountability_The_impact_of_instructional_assessment_on_elementary_children%27s_achievement_test_scores**

Methe, S. A., Begeny, J. C., & Leary, L. L. (2011). Development of conceptually focused early numeracy skill indicators. *Assessment for Effective Intervention*, *36*(4), 230–242. **https://doi.org/10.1177/1534508411414150**

Miller, D., & Lavin, F. (2007). But now I feel I want to give it a try: Formative assessment, self-esteem, and a sense of competence. *The Curriculum Journal, 18*(1), 3–25. **https://doi.org/10.1080/09585170701292109**

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Murata, A. (2004). Paths to learning ten-structured understanding of teen sums: Addition solution methods of Japanese Grade 1 students. *Cognition and Instruction*, *22*, 185–218.

Murata, A., & Fuson, K. C. (2006). Teaching as assisting individual constructive paths within an interdependent class learning zone: Japanese first graders learning to add using 10. *Journal for Research in Mathematics Education, 37*(5), 421–456. **https://storage.googleapis.com/wzukusers/user-33451779/documents/66b4407bf52a4a2f9d4a6d57f52cd8aa/23.%20Individual%20constructive%20paths%20within%20an%20interdependent%20CLZ.pdf**

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics.*

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. U.S. Department of Education.

National Research Council. (2001a). *Adding it up: Helping children learn mathematics* (J. Kilpatrick, J. Swafford, & B. Findell, Eds.). National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

National Research Council. (2011). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

NGSS Lead States. (2013). Next generation science standards for states, by states. National Academies Press. **https://doi.org/10.17226/18290.**

Ozan, C., & Kıncal, R. Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences: Theory & Practice, 18,* 85–118. **http://dx.doi.org/10.12738/estp.2018.1.0216**

Paik, S., Song, G., Kim, S., & Ha, M. (2017). Developing a four-level learning progression and assessment for the concept of buoyancy. *Eurasia Journal of Mathematics, Science and Technology Education, 13(8)*, 4965–4986.

Pearson. (2005). *Achieving student progress with scientifically based formative assessment: A white paper from Pearson*.

Pellegrino, J. W. (2005). The challenge of knowing what students know. *Measurement: Interdisciplinary Research and Perspectives*, *1*(2). e7–e11.

Pellegrino, J. W. (2018). Assessment of and for learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 410–421). Routledge.

Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1*(2), 103–148.

Pellegrino, J. W., & Hilton, M. (Eds.) (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* National Academies Press.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the next generation science standards.* National Academies Press.

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. **https://doi.org/10.1080/00461520.2016.1145550**

Penuel, W. R. (2015). Learning progressions as evolving tools in joint enterprises for educational improvement. *Measurement: Interdisciplinary Research and Perspectives*, *13*(2), 123–127.

Piaget, J. (1964). Development and learning. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget rediscovered* (pp. 7–20). Cornell University.

Platas, L. M., Ketterlin-Geller, L. R., & Sitabkhan, Y. (2016). Using an assessment of early mathematical knowledge and skills to inform policy and practice: Examples from the early grade mathematics assessment. *International Journal of Education in Mathematics, Science and Technology*, *4*(3), 163–173. **https://eric.ed.gov/?id=EJ1102189**

Popham, W. J. (2008). *Transformative assessment*. Association for Supervision and Curriculum Development.

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology, 105*(2)*,* 453–464. **https://doi.org/10.1037/a0031753**

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction*, *60*, 154–165.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for instruction*. Erlbaum. Rich, K. M., Binkowski, T. A., Strickland, C., & Franklin, D. (2018). Decomposition: A K-8 computational thinking learning trajectory. *In ICER '18: Proceedings of the 2018 ACM Conference International Computing Education Research*. ACM. **https://doi.org/10.1145/3230977.3230979**

Rittle-Johnson, B., Fyfe, E. R., Loehr, A. L., & Miller, M. R. (2015). Beyond numeracy in preschool: Adding patterns to the equation. *Early Childhood Research Quarterly*, *31*, 101–112. **https://doi.org/10.1016/j.ecresq.2015.01.005**

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, *17*(1), 1–24.

Ruiz-Primo, A. M., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning, *Educational Assessment*, *11*(3 & 4), 205–235.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R.M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Shea, N. A., & Duncan, R. G. (2013). From theory to data: The process of refining learning progressions. *Journal of the Learning Sciences*, *22*(1), 7–32.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018a). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Classroom assessment principles to support learning and avoid the harms of testing. *Educational Measurement: Issues and Practice, 37*(1), 52–57. **https://doi.org/10.1111/emip.12195**

Shinn, M. R. (1995). Best practices in curriculum-based measurement and its use in a problem solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (3rd ed., pp. 547–567). National Association of School Psychologists.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145. **https://doi.org/10.2307/749205**

Smith, C., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and atomic-molecular theory. *Measurement*, *14*(1&2), 1–98.

Steedle, J. T. (2006, April). *Seeking evidence supporting assumptions underlying the measurement of progress variable levels*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Stiggins, R. (2002). Assessment Crisis: The absence of assessment for learning. *Phi Delta Kappan*, June, 758–765.

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction: Toward a theory of teaching. *Educational Researcher*, *41*(5), 147–156. **https://doi.org/10.3102/0013189X12442801**

Van den Heuvel-Panhuizen, M. H. A. M. (1996). *Assessment and realistic mathematics education* (Vol. 19). Utrecht University.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wearne, D., & Hiebert, J. (1988). A cognitive approach to meaningful mathematics instruction: Testing a local theory using decimal numbers. *Journal for Research in Mathematics Education*, *19*(5), 371–384.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2006). *Systems for state science assessment*. National Academies Press.

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. **https://doi.org/10.1016/j.jmathb.2012.12.003**

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 359–403). Sense Publishers.

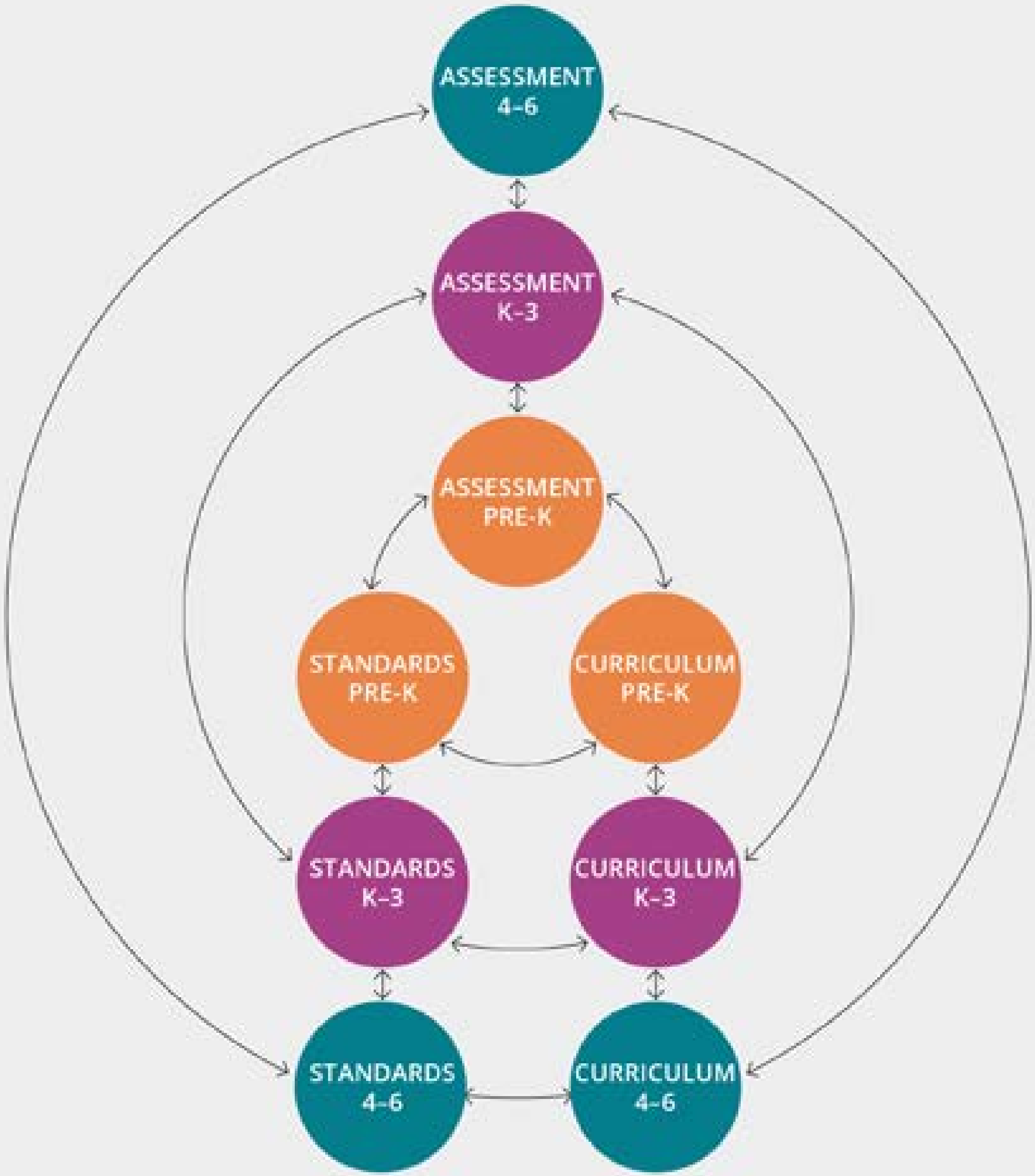
Zaidi, S. Z., Arnold, S., Lehman, E. M., & Strickland, C. (2022). Beyond the design of assessment tasks: Expanding the assessment toolkit to support teachers' formative assessment practices in elementary science classrooms. In C. Chinn, E. Tan, C. Chan, & Y. Kali (Eds.), *International collaboration toward educational innovation for all: Overarching research, development, and practices, Proceedings of the 16th International Conference of the Learning Sciences (ICLS) 2022,* 1964-1965. International Society of the Learning Sciences.

Zhang, L., & Nouri, J. (2019). A systematic review of learning computational thinking through Scratch in K–9. *Computers & Education*, *141*, 103607 **https://doi.org/10.1016/j.compedu.2019.103607**

Suggested Citation

Baroody, A., & Pellegrino, J. W. (2023). Assessment for learning. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives* (pp. 83–94). Community for Advancing Discovery Research in Education (CADRE), Education Development Center, Inc.

Figure 2. Alignment of assessment vertically as well as laterally with goals or standards and curriculum. (The symbol ⟷ indicates an alignment between two elements; graphic used with permission of Arthur Baroody, author.)



Note. Middle school and high school levels are not shown because of space limitations

Table 1. An example of learning progression (based upon Baroody & Purpura, 2017).

Level in learning progression	Relation to achieving fluency with the basic sums and differences
E1. **Verbal subitizing (verbal-based cardinal concept of small numbers; 1 and 2 first, then 3, and—in time—4 to 6)**	Most fundamentally, verbal subitizing provides a basis for understanding the number words used in arithmetic word problems and the numerals used in written arithmetic expression. It enables children to discover important properties about small numbers and operations on these "intuitive" numbers (Levels R1, R2, R5 and O1 to O3, in particular), properties that can be generalized to all counting or natural numbers (Sarnecka & Gelman, 2004).
R1. **Pre-counting relative magnitude of small collections (ordinal meaning of number words)**	R1 is discovered by literally seeing (via subitizing) that "three" is more than "two" and "two" is more than "one." R1 provides a basis for a local understanding of the increasing-magnitude principle (the insight that the order of the number words for small numbers—"one, two, three"—reflects increasing magnitude) and inducing local and then general knowledge of the stable order principle (the understanding of why we say "one, two, three" in order and later why saying number words in order matters). This, in turn, provides a basis for the generalization that the counting sequence represents increasingly larger quantities (Levels R2 to R5) and a tool for mentally representing the effect of addition and subtraction (Level O1) and informally computing sums and differences (Levels O2 and O3).
O1. **Phase 0: Basic addition and subtraction concepts and nonverbal addition and subtraction**	Verbal subitizing provides a basis for representing small collections and mentally operating on them—success on a nonverbal addition/subtraction task. Such a task involves showing a child a collection on a mat, covering it, putting out additional items and then putting them under the cover or removing and showing one or more hidden items, and finally, asking a child to make his/her mat like the tester's. Concrete nonverbal addition (with visible collections): Verbally subitizes the original collection, then the amount added and finally the total (e.g., sees 2 cookies, sees 1 more added; sees total is "more"—specifically "three" [composition]. Additionally, sees a collection three as composed of two items and one item [decomposition]). Abstract nonverbal addition (with mental representations): Concrete nonverbal addition experiences lead to the construction of an *informal change add-to concept of addition* (adding items to a collection makes the original collection larger) and the ability to mentally represent collections and determine the sum. These developments provide a foundation for counting-based addition and subtraction (Level O2).
C1. **Verbal counting—string level and** C2. **Verbal counting—unbreakable chain level**	Level C1 (e.g., "onetwothree") provides a basis for Level C2 (e.g., "one, two, three"), which in turn, provides a basis for Level C3 (e.g., the ability to start counting at "two") and its applications to adding and subtracting (Levels O2 and O3).

Table 1. Continued

Level in learning progression	Relation to achieving fluency with the basic sums and differences
E2. **Meaningful object counting**	Subitizing experiences can help children construct the following counting principles that underlie *meaningful object counting: stable order* (number words must be recited in a standard order), *abstraction* (a collection can be composed of different-looking items), *one-to-one* (each item in collection should be tagged with one and only one number word), *cardinality* (the last number word represents the total), and *order-irrelevance principles* (the order of one-to-one counting does not affect the total; see Baroody, Lai, & Mix, 2006, for details). Meaningful object counting is necessary for children to devise counting-based addition and subtraction strategies (Levels O2 and O3).
R2. **Increasing magnitude principle plus counting-based number comparisons (especially collections larger than 3)**	Recognizing "the counting sequence represents increasingly larger quantities enables children to use *meaningful object counting* to determine the larger of two collections (e.g., 7 items is more than 6 items because you have to count further to get to *seven* than you do for *six*)" (Baroody, 2016). This may lay the foundation for Levels R3 to R5 and O2.
R3. **Mental comparisons of non-neighboring or non-successive numbers**	*Familiarity with the counting sequence* and *the increasing magnitude concept* provide a basis for comparing two numbers at obviously different positions in the counting sequence (i.e., to make gross comparisons of 2 or 7, 10 or 3, 9 or 5, and 4 or 8) (Baroody, 2017).
C3. **Verbal counting—breakable chain level (number-after knowledge)**	Familiarity with the counting sequence enables a child to enter the sequence at any point *and specify the next number* instead of always counting from one (Baroody, 2017).
R4. **Mental comparisons of neighboring or successive numbers (number after equals more)**	The use of the *increasing magnitude principle* and *number-after knowledge* enables children to determine efficiently and *mentally compare even close numbers* such as the larger of two neighboring numbers (e.g., immediately responding "eight" to the question: "Which is more seven or eight? because eight follows seven in the counting sequence) (Baroody, 2017).
R5. **Successor principle (number after equals 1 more) and reconceptualization of the counting sequence as the (positive) integer sequence**	Verbal subitizing enables children to see that "two" is exactly one more than "one" item and that "three" is exactly one more than "two" items, and this can help them understand the *successor principle* (each successive number in the counting sequence is exactly one more than the previous number). The *successor principle* enables children to view the counting sequence as n, $n + 1$, $[n + 1] + 1$, . . . (*positive integer sequence*)—a linear representation of number (Baroody, 2017).
O2. **Basic Phase 1: Concrete counting-based strategies for determining sums or differences**	An *informal change add-to concept of addition* enables direct modeling of the change add-to concept of addition by representing (one or both) addends done first and then determining the sum (Baroody, 2017).

Table 1. Continued

Level in learning progression	Relation to achieving fluency with the basic sums and differences
C5. **Verbal counting—numerable chain level**	Recognition that the counting words themselves can be counted provides the basis for Level O3.
O3. **Advanced Phase 1: Abstract counting-based strategies for determining sums or differences**	Indirect modeling of children's informal change add-to view of addition is done by representing both addends and determining the sum *simultaneously.*
O4. **Phase 2: Deliberate rule- and reasoning-based strategies for determining sums**	Counting-based (Level O2 and especially relatively efficient Level O3) strategies provide an important basis for discovering patterns and relations that serve as the basis for rules or reasoning strategies. Easiest combinations: The most salient patterns or relations are the add-0 rule and number-after *n* rule for +1 (e.g., 4 + 1 is the number after "four" when we count: five), which requires an extension of the *breakable chain level*, namely *fluency with number-after relations*). Relatively difficult combinations: Other deliberate reasoning strategies that involve using what is known to deduce the sum).
O5. **Phase 3: Automatic and fluent retrieval from a memory network**	The retrieval network may be composed of an interconnected web of facts and relations and involve both reproductive (fact recall) and reconstructive (automatic rule-or reasoning-based) strategies.

Equity and Justice in Classroom Assessment of STEM Learning

Erin Marie Furtak, University of Colorado Boulder, and
Okhee Lee, New York University

3

Assessment has the potential to help us understand what needs to be transformed and help us support a kind of learning that can cultivate just, sustainable and culturally thriving communities.

—Megan Bang, 2019

Assessments of student learning in school subjects, and in STEM subjects specifically, have traditionally been conceived as neutral, although they are not (Taylor, 2022). Which constructs are valued and how systems privilege canonical and dominant ways of knowing influence how students perform and who succeeds (Au, 2007; Randall, 2021). These larger realities set boundaries around what is possible and relevant to classroom assessment of STEM learning (Nasir et al., 2020; Noble et al., 2012; Wright & Riley, 2021).

At present, as the ways we understand and theorize learning have expanded to sociocultural frames (NASEM, 2018b), reforms in STEM learning have increasingly focused on engagement in disciplinary practices (Ford & Forman, 2006). For example, students engage in science and engineering by blending practices, crosscutting concepts, and disciplinary core ideas (NGSS Lead States, 2013a). Similar reforms have taken hold in mathematics education by focusing on mathematical practices (NGA Center for Best Practices & CCSSO, 2010) and in computer science (CS) education by focusing on computational practices (K–12 Computer Science Framework, 2016).

At the same time, as U.S. schools become increasingly diverse, our design of learning environments and classroom assessments seeks to create better connections with students' interests, identities, and cultures (NRC, 2018). Reforms in STEM learning are aimed at equity by providing access and promoting achievement (e.g., NRC, 2014 in science education). This vision encourages designers of learning experiences to link STEM learning to relevant phenomena and problems (NRC, 2012) and to foster more inclusive classroom cultures (NGA Center for Best Practices & CCSSO, 2010; K–12 Computer Science Framework, 2016). This vision is also explicitly intended to promote equitable engagement and broaden opportunities for youth from minoritized groups, including students in poverty, students of color, students with disabilities, and multilingual learners. For example, the NGSS are intended for all students, hence "all standards, all students" (NGSS Lead States, 2013b), and the K–12 Computer Science Framework (2016) highlights equity in computer science education (**https://k12cs.org/wp-content/uploads/2016/09/K%E2%80%9312-Computer-Science-Framework.pdf.**).

It has now been about a decade since standards-based reforms in mathematics education (NGA Center for Best Practices & CCSSO, 2010) and science education (NGSS Lead States, 2013a; NRC, 2012; 2014) compelled significant shifts in curriculum, instruction, and assessment in STEM learning environments. Significant progress has been attained in the design of curriculum materials (Campbell & Lee, 2021; Edelson et al., 2021) and ambitious instructional practices to support student engagement with these materials (Shepard, 2021). However, assessment is still catching up to the vision of current reforms.

The call to reframe STEM learning experiences is even more urgent with pressing societal challenges (e.g., COVID-19 pandemic, climate change) that disproportionately impact minoritized groups and further expose historical and present systemic injustices. As students make sense of pressing societal challenges, they are better equipped to make informed decisions about these challenges, take responsible actions, and design solutions for a more just society (Grapin et al., in press; Lee & Grapin, 2022). Moreover, assessment could address how multiple ways of knowing can inform STEM learning, how assessment can be designed to sustain learners' cultural practices and knowledge, and how assessment can be designed in partnership with learners, families, and communities (Tzou et al., 2021).

In this section, we first discuss conceptions of equity and justice in the emerging literature in STEM learning. Then, we describe how classroom assessment intersects with curriculum and instruction in STEM subjects. Next, we draw upon DRK-12 projects to illustrate how equity and justice are addressed in classroom assessment of STEM learning. Finally, we offer future directions for equity and justice in STEM classroom assessment.

Conceptions of Equity and Justice in Classroom Assessment of STEM Learning

The DRK-12 program is designed to "significantly enhance the learning and teaching of science, technology, engineering, mathematics and computer science (STEM) by pre-K–12 students and teachers through research and development of STEM education innovations and approaches" (National Science Foundation [NSF], 2020). The program has three major research and development strands: assessment, learning, and teaching, which it acknowledges are both overlapping and independent of each other.

The DRK-12 program "seeks to maintain a balanced portfolio" (NSF, 2020) by supporting projects both to meet the current needs in STEM education (i.e., contemporary approaches) and to challenge existing assumptions and offer promising new approaches to STEM learning, teaching, and assessment (i.e., future approaches), as described in the text of the DRK-12 solicitation below (NSF, 2020.):

The DRK-12 program seeks to maintain a balanced portfolio by supporting projects ranging from those with immediate applicability to those that anticipate and provide the foundation for pre-K–12 STEM education as it could be in future decades. The DRK-12 program is interested in ideas, concepts, theories, practices, and research and development that can challenge existing assumptions and offer promising new approaches to STEM learning, teaching, and assessment. Such proposals could, for example, adapt research-based effective STEM teaching practices in ways that explicitly support the broader and increasingly diverse set of learners who comprise the nation's student population. (p. 4)

As stated above, the DRK-12 program funds projects that "challenge existing assumptions and offer promising new approaches" that could also "explicitly support the broader and increasingly diverse set of learners who comprise the nation's student population" (NSF, 2020). The term *diversity* as used in the literature indicates multiple forms of diversity, including gender diversity (Gunckel, 2019); racial diversity (Basile & López, 2015; Delgado & Stefancic, 2006); linguistic diversity (Lee et al., 2013; NASEM, 2018a); economic diversity (Calabrese Barton, 1998; McKinney de Royston & Farinde, 2018); and diversity of disabilities (Guzman-Orth et al., 2021).

For these multiple forms of diversity, conceptions of equity, justice, and related terms are not straightforward. Philip and Azevedo (2017) argued that in STEM learning environments, "implicit and explicit values and goals … are intertwined with conceptions of equity" (p. 527). They presented four approaches to equity: "(1) increasing opportunity and access to high-quality science and engineering learning and instruction; (2) emphasizing increased achievement, representation, and identification with science and engineering; (3) expanding what constitutes science and engineering; and (4) seeing science and engineering as part of justice movements" (cited in NASEM, 2022, pp. 23–27).

Building on Philip and Azevedo's four approaches, the NASEM consensus report *Science and Engineering in Preschool Through Elementary Grades: The Brilliance of Children and the Strengths of Educators* (2022) provided the following definitions of "equity" and "justice":

This report uses the term "equity" to address ways—through changing policies and practices—to remove barriers to participation in science and engineering and increase achievement, representation, and identification (mainly the first two approaches, though all four approaches work toward equity). Equity thus strives for comparable levels of attainment and/or participation. The report uses the term "justice" to refer specifically to addressing systemic oppressions that cause those barriers (mainly the third (p. 23) . . . and fourth approaches), seeking fair treatment of all people and supporting opportunities for self-determination and thriving. (p. 27)

We expand these definitions of equity and justice to articulate how they can be enacted in classroom assessment of STEM learning. When we consider *equity* (the first two approaches in Philip & Azevedo, 2017), we consider how assessment policies and practices can increase access (approach 1) and achievement (approach 2). Equity in assessment of STEM learning includes assessment designs that create more expansive space for learners to show what they know and are able to do (Fine & Furtak, 2020; Kang et al., 2014; Lee et al., 2019); assessments contextualized with phenomena and authentic problems relevant to diverse learners' lives (Edelson et al., 2021; Penuel et al., 2019); assessments that are accessible for students with disabilities and neurodiversity (Rose et al., 2018); and assessments that allow and encourage translanguaging by multilingual learners (Fine, 2022).

When we consider *justice* (the last two approaches in Philip & Azevedo, 2017), we consider how assessments can redefine what counts as STEM learning (approach 3) and how to reframe STEM learning toward a just society (approach 4). Justice in STEM learning challenges the dominance of White Western worldviews in science and STEM disciplines (Bang & Medin, 2010; Harding, 2008; Warren et al., 2020; Randall, 2021) and promotes how multiple ways of knowing can inform STEM learning (Kimmerer, 2013; Tzou et al., 2019). These approaches include designing curriculum and assessment in partnership with learners, families, and communities (Tzou et al., 2019; 2021); seeking ways to design classroom assessments that sustain learners' cultural practices and knowledge (Burgess & Patterson Williams, 2022; Likely, 2020; Randall, 2021); and repositioning classroom assessment as a way of transforming learners' experiences in STEM (Kang et al., 2022). These approaches also connect opportunities for students to show their knowledge and abilities to larger societal movements, such as those for climate justice (Leckey et al., 2021), and to sustaining practices from their homes and communities (Wright & Riley, 2021).

Teachers need professional learning opportunities in order to promote equity and justice in STEM classroom assessment. Equity in STEM classroom assessment includes professional learning opportunities that help teachers attend to and be more responsive to students' ideas in the course of everyday, informal assessment settings (Kang, 2022) as well as student disabilities and neurodiversity (Israel et al., 2022). Justice in classroom assessment of STEM learning includes frameworks to guide teachers as they interpret students' and families' science ideas (Learning in Places Collaborative, 2020) and experiences of place (Semken et al., 2017) as well as explorations of teachers' own identities (Taylor, 2022).

Equity and Justice in Classroom Assessment of STEM Learning

Equity and justice are not only at the core of STEM classroom assessment, but part of learning environments that link assessment with curriculum and instruction (**Figure 3**). This approach highlights that separating classroom assessment from learning environments is not possible (Kang & Furtak, 2021), and effective approaches to the design and use of classroom assessments are supported by curriculum and instruction (Shepard et al., 2018). Building on learning theories as guides for learning goals, instruction, and assessment (see **Section 2** of this report), we argue that equity and justice should also be placed at the center of STEM learning environments.

Figure 3. Intersections of curriculum, instruction, and assessment centering equity and justice (graphic used with permission of Erin M. Furak, author)

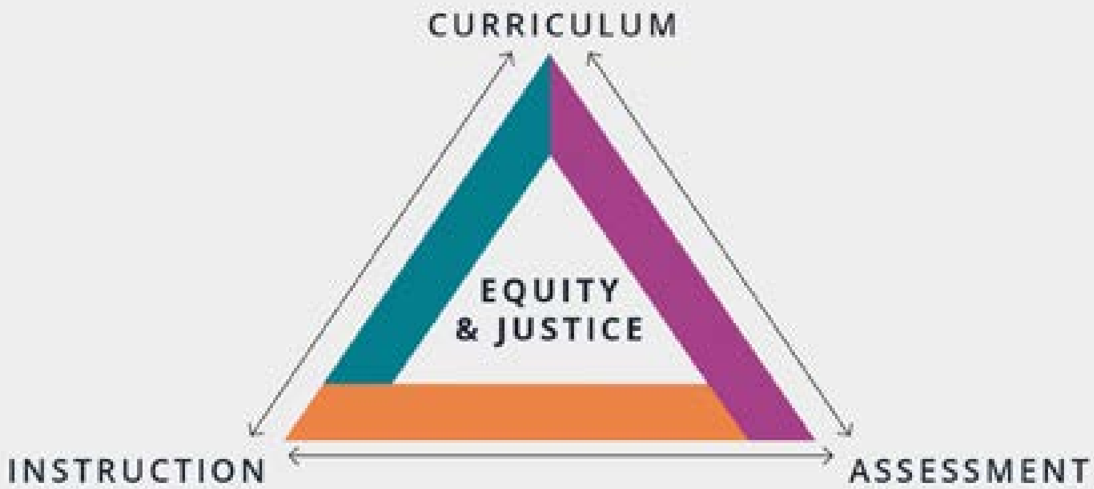


Figure 3 illustrates intersections of curriculum, instruction, and assessment centering equity and justice in STEM learning (NASEM, 2022). This triangle helps us to look beyond past approaches to classroom assessment of STEM learning: once core elements were designed for "all," and accommodations or modifications as "adds-on" were provided to guide teachers to adapt curriculum, instruction, and assessment for specific student groups. These past approaches sought to broaden participation of diverse student groups in STEM learning (Philip & Azevedo's [2017] approaches 1 and 2) but did not address existing power structures and hierarchies and, as a result, did not disrupt dominant paradigms and epistemologies (Philip & Azevedo's [2017] approaches 3 and 4; Calabrese Barton et al., 2020; NASEM, 2022). For equity and justice to be centered in classroom assessment of STEM learning, it must be accompanied by overall transformations in curriculum and instruction.

Curriculum. Centering equity and justice begins with curriculum materials centered on phenomena and problems that are authentic and relevant to learners' lives, as well as includes mechanisms for neurodiverse learners to access and engage with phenomena and problems. Curriculum materials designed to promote equity and justice are developed in partnership with teachers, students, and communities using phenomena and problems located in place. These curriculum materials also embed multiple and ongoing opportunities for learners to share their experiences with families and communities.

Instruction. Enactment of everyday learning experiences in reform-oriented STEM learning environments involves multiple opportunities for classroom assessment, in which teachers create expansive space for students to share their thinking and ways in which they demonstrate their learning, and for teachers to be responsive to their ideas (Kang, 2022). Professional learning prepares teachers to incorporate equity and justice-oriented classroom practices and also includes space for teachers to reflect on their own experiences, histories, and connections to place.

Assessment. As described above, assessment that centers equity and justice goes beyond seeking to narrow perceived achievement gaps and broadening access and opportunities for STEM learning toward transformative approaches that are based in communities and places. Both assessment tasks and teachers' everyday assessment practices can build on students' interests and assets, acknowledge multiple ways of knowing, and create space for students' engagement and communication.

Examples from the DRK-12 Portfolio

We provide examples of DRK-12 projects funded in recent years in different STEM subjects. Across these projects, we articulate how they conceptualize equity and justice at intersections of curriculum, instruction, and assessment. For each project, we describe its purpose, how assessment intersects with curriculum and instruction, and how assessment centers equity and justice.

Computer Science Education

A Research-Practice Partnership for Developing Computational Thinking through Linguistically and Culturally Relevant CS Curriculum in Middle School (Gray & Escandon, NSF 1923586, collaborative with Mortimer & Akbar, NSF 1923599, 2019-2023)

Purpose. The *Sol y agua*[2] project explores how bilingual students' knowledge of borderlands ecology can develop their computational thinking as they approach and solve questions of relevance to them. The project involves both local scientists and Native Americans' stewardship in its processes of designing a pilot game. The game is embedded within a bilingual (Spanish/English) middle school curriculum module that supports translanguaging by allowing students to switch between English and Spanish seamlessly.

Intersections among curriculum, instruction, and assessment. Within the *Sol y agua* curriculum, students are empowered to select questions and issues of relevance to them, and community members are involved in providing feedback during ongoing processes of design. The project promotes teachers to develop knowledge of translanguaging, CS, and culturally relevant pedagogies to support students as they engage with the project-based module.

Equity and justice in assessment. The project centers equity in its assessment designs that broaden students' opportunities to demonstrate what they know and are able to do by leveraging both students' knowledge of the Chihuahuan desert and engagement in computational knowledge and practices. The assessment designs also create space for students to engage in translanguaging.

2 See Further Reading at the end of this section for more information on the *Sol y agua* project.

Engineering Education

Community-Based Engineering Design Challenges for Adolescent English Learners (Wilson-Lopez, Hailey, & Householder, NSF 1222566, 2012-2016)

Purpose. This project was an exploratory ethnographic study conducted in partnership with Latino/Latina communities. The approach was based on a theoretical model that views engineering design practices as embedded within social, cultural, and linguistic activities. By following a group of high school students as they solved engineering design challenges relevant to their own communities, the project positioned learners' resources and linguistic practices as central to their approaches to engineering design.

Intersections among curriculum, instruction, and assessment. Learners were supported in after-school settings as they identified engineering design challenges of relevance to them. They worked in groups and were supported by members of the research team in drawing upon their cultural and linguistic resources to solve the problems they identified. The research team asked students at specific points where they were in the engineering design process and to reflect on their own work (Wilson-Lopez et al., 2016). Students' design projects included wheelchair-accessible doors, headrests in shower chairs, a rainwater catchment system, and improvement of a community playground.

Equity and justice in assessment. The project promoted justice in assessment by inviting learners to identify engineering problems relevant to their lives and communities and by creating space for learners to leverage multiple forms of capital for engineering designs. At the conclusion of their designs, learners were provided with opportunities to share their designs with possible funders. The project highlights how the learners and their communities brought diverse forms of cultural and social capital that strengthened their engagement in engineering designs (Wilson-Lopez et al., 2018).

Mathematics Education

Collaborative Research: Advancing Equity and Strengthening Teaching with Elementary Mathematical Modeling (Aguirre, Carlson, Suhg, & Turner, NSF 2010269, 2008997, 2010202 & 2010178, 2020-2024)

Purpose. The project explores how to rehumanize elementary mathematics (Gutiérrez, 2018) through equity-centered mathematical modeling. The project seeks to broaden access and opportunity through creating modeling tasks and routines that make connections with students' cultural and community contexts. Project activities are built on an approach for how culturally responsive mathematics teaching can promote equity, including a focus on cultural and community funds of knowledge, affirming multilingualism, disrupting power and status, and analyzing and taking action for justice.

Intersections among curriculum, instruction, and assessment. The project[3] has created modules that provide opportunities for elementary mathematics teachers to learn about

3 See Further Reading at the end of this section for more information on *Advancing Equity and Strengthening Teaching with Elementary Mathematical Modeling.*

mathematical modeling, as well as how social issues can create critical contexts for teachers to support students as they engage with mathematical models. Modules for teacher learning focus on instructional practices to support student engagement, including fostering creativity, seeing strengths in students' model-building, and responding to student ideas and fostering growth.

Equity and justice in assessment. The project integrates formal and informal assessment opportunities that promote equity, including modules (described above) that support informal assessment practices, such as listening and responding to student ideas and experiences, as well as instructional strategies that support students' ongoing growth in mathematical modeling. The project has also created multiple mathematical modeling tasks which can serve as opportunities for assessment that integrate structures that center student experiences and ideas, such as the amount of waste produced in a school cafeteria or how many jugs of water a class needs to have safe water for a whole day. These tasks are implemented with routines for civic awareness and action as an integral part of mathematical modeling for justice (Aguirre et al., 2019).

Science Education

Preparing Teachers to Design Tasks to Support, Engage, and Assess Science Learning in Rural Schools (Penuel, Lo, & Wingert, NSF 2010086, 2020-2024)

Purpose. This project partners with teachers in rural high schools to create professional learning opportunities around enactment of the three-dimensional (3D) learning that is the foundation of the *Framework* (NRC, 2012) and the NGSS. Then, the project adds two additional dimensions—student interest and identity—as design principles for "5D assessment."

Intersections among curriculum, instruction, and assessment. The project creates an online professional learning space to support teachers in learning how to connect instruction and assessment to students' everyday lives. Teachers participate in a multi-step process in which they (a) unpack the NGSS relevant to their curriculum, (b) learn about students' interests and experiences relevant to what will be taught, (c) learn about the role of phenomena relevant to the NGSS, (d) select phenomena related to what they learned about students, and (e) design assessments that promote students' engagement in 5D learning. This integrated approach creates space for teachers to locate student interest and identity at intersections of curriculum, instruction, and assessment (Lo et al., 2022). Project learning experiences create space for teachers to learn about phenomena and how phenomena can serve as the target of explanations for science assessments.

Equity and justice in assessment. Although using phenomena and problems relevant to diverse student groups to promote equity has been widely accepted in science curriculum and instruction, phenomenon-based science assessment is a relatively new approach. Moreover, by naming students' interests and identity as equal partners with the other three dimensions of the NGSS, this project promotes equity as part of the foundational design of assessments.

Future Directions for Classroom Assessment of STEM Learning

Based on the *Framework* (NRC, 2012) and the NGSS (NGSS Lead States. (2013a), the NGSS assessment report (NRC, 2014) offered the following recommendations on the design of assessments and assessment systems (see Chapter 7 of that report):

1. **Consider multiple dimensions of diversity**, including culture, language, ethnicity, gender, and disability;
2. **Collect relevant indicators about opportunity to learn**—including material, human, and social resources available to support student learning—to contextualize and validate the inferences drawn from the assessment results;
3. **Value and respect** the experiences that all students bring from their backgrounds (e.g., homes and communities);
4. **Articulate students' background knowledge** (e.g., cultural or linguistic knowledge) with disciplinary knowledge;
5. **Offer sufficient school resources** to support student learning; and
6. **Support teacher professional development**.

The NRC's (2014) recommendations address equity in science assessment substantially, as it is represented by three of the six recommendations (1, 3, and 4). The recommendations for equity and justice were forward thinking, considering that the NGSS assessment report (NRC, 2014) was released almost a decade ago. Understandably, these recommendations are more consistent with seeking to broaden participation in STEM—that is, associated with equity approaches—as opposed to disrupting existing power dynamics, as in approaches associated with justice.

Consistent with the NRC's (2014) recommendations, the DRK-12 solicitation asks applicants—as it has for several years—to locate their research within one of three strands: assessment, learning (including STEM curriculum materials), and teaching (including teacher preparation and professional development). The purpose of equity in assessment is highlighted in the overall description for all three strands: "STEM assessment innovations or approaches must be fair and equitable, as well as culturally and linguistically sensitive in addressing the needs of the learners under study" (NSF, 2020, p. 5). However, equity is not included as possible research questions within the assessment strand, whereas specific examples of equity as possible research questions are provided within the learning and the teaching strands (NSF, 2020, pp. 5-6).

In the absence of the consideration of equity in the assessment strand of the DRK-12 program, we looked into emerging literature on equity and justice. We use the definitions of equity and justice in the recent NASEM report (2022), which is based on four approaches to equity by Philip and Azevedo (2017). Our searches of the projects listed as "assessment" in the DRK-12 portfolio indicate that many projects include accommodations for diverse student groups, including projects that explore assessment designs for students with disabilities.[4] A small but growing number of DRK-12 assessment projects take approaches centered on equity, and a few DRK-12 assessment projects take approaches centered on justice. Those projects that center on equity and/or justice in assessment of STEM learning do so as classroom assessment intersects with curriculum and instruction. In addition, these projects design classroom assessment tasks and practices in partnership with learners, families, and communities. This work is deliberately expansive to include the many places where assessment occurs in STEM learning environments: when we design instructional materials, as we prepare teachers through professional learning, during everyday teaching and learning interactions, at the conclusion of learning experiences, and working with families and communities beyond school settings.

Based on our analysis of DRK-12 assessment projects grounded in the NRC's (2014) recommendations for assessment and the emerging literature on equity and justice (NASEM, 2022; Philip & Azevedo, 2017), we make three recommendations on equity and justice in assessment of STEM learning—one for the NSF, one for research, and one for practice.

First, we encourage the NSF to create more expansive space for funded projects at intersections of curriculum, instruction, and assessment. Assessment of STEM learning for the purposes of equity and justice is difficult to locate within only one of these strands. As the examples in this chapter have illustrated, projects that promote equity and justice in assessment do so in concert with curriculum materials and through support of instruction and professional learning. We need to better understand students' experiences with assessment, their opportunities to engage in meaningful learning, the ways they are supported through instruction, and how their ideas and experiences are represented in and taken up in learning. Such understanding is necessary to contextualize what assessments can tell us about students' knowledge and abilities. We also need to look toward models that seek to design assessments in partnership with learners, parents, and community members (e.g., Randall, 2021).

Second, we encourage the research community to take an expansive view of equity and justice. Our analysis of the NSF DRK-12 portfolio indicates that more of the projects are taking approaches associated with equity than justice in assessment. This is to be expected, as current approaches to assessment of STEM learning are based on the current standards in mathematics education (CCSS-M, 2010) and science and engineering education (NRC, 2012; NGSS Lead States, 2013a). These projects seek to identify design features that make assessments more accessible to learners (e.g., using wider ranges of phenomena, using multimodality) to create more space for learners to share what they know and be able to do. Expanding this literature, we advocate for moving toward justice in classroom assessment of STEM learning (Philip & Azevedo, 2017). This approach destabilizes "settled" expectations and criteria for what counts as science and is assessed in science (Bang et al., 2013); whose knowledge and values are represented in the questions asked on the assessments; how assessments acknowledge and integrate multiple ways of knowing; and how designers build assessments *with*, and not *for,* learners and communities (Bang, 2019). This approach also disrupts historical power structures in STEM by

4 See for example, Fraction Activities and Assessments for Conceptual Teaching [FAACT] for Students with Learning Disabilities; Jessica Hunt, PI; NSF 1708327. **https://www.nsf.gov/awardsearch/showAward?AWD_ID=1708327**

creating learning environments where students are thriving (Bang, 2019) and re-humanized (Gutiérrez, 2018), toward Indigenous resurgence (Tzou et al., 2019), toward antiracism (Randall, 2021), and for Black Joy (Love, 2019). Designing assessments that center justice and look toward new futures in STEM learning will require fundamental shifts in the field by broadening what counts as science (Bang et al., 2018; Tzou et al., 2021), exploring how students design justice-centered solutions to pressing societal challenges that disproportionately impact minoritized groups (Grapin et al, in press; Lee & Grapin, 2022), and expanding "who is at the table" when assessments of STEM learning are designed.

Finally, we encourage the practice community to embrace their embedded roles in communities. Teachers, school leaders, and school districts are ideally located to center equity and justice in developing or adapting curriculum materials and assessments and in implementing instructional approaches that build on their relationships with students, families, and communities. In particular, teachers need spaces to reflect upon the role that their classroom assessment plays in students' opportunities to engage in STEM learning. Professional learning opportunities can provide support for teachers to engage with and learn about the interests and problems encountered within the communities where they work, and to use community interests and problems as foundations for classroom assessments (e.g., Lo et al, 2022).

Box 3: Example from the DRK-12 Portfolio: Science and Integrated Language [NSF 1503330]

Scott Grapin, University of Miami

Science And Integrated Language (*SAIL*) is a yearlong fifth-grade science curriculum with a focus on promoting both science learning and language learning with multilingual learners. Grounded in design principles that capitalize on mutually supportive instructional shifts in the fields of science education and language education (Lee et al., 2019), *SAIL* consists of four 9-week units that address all 16 NGSS performance expectations for fifth-grade. Over a four-year period, *SAIL* underwent iterative cycles of development, testing, and revision with teachers and students in linguistically diverse science classrooms. The curriculum includes a suite of assessments, including four types of formative assessments embedded throughout the units and a summative assessment at the end of each unit (Llosa et al., 2022). *SAIL* is accompanied by a professional development program that engages teachers in asset-oriented, collaborative learning that is symmetrical (Mehta & Fine, 2019) to the kinds of learning experiences that teachers create for their students in the curriculum.

SAIL instruction centers equity for multilingual learners, a group historically denied meaningful science learning experiences (e.g., NASEM, 2018a). *SAIL* provides scaffolded opportunities for multilingual learners to make sense of locally relevant science phenomena as they engage in goal-directed interactions with their peers and teacher in a supportive community of practice (NASEM, 2022). In addition, *SAIL* instruction invites students to draw from their full repertoire of meaning-making resources, including resources that have not always been valued or legitimized in science instruction with multilingual learners, such as nonlinguistic modalities (beyond a narrow focus on the linguistic modality) and everyday registers (beyond a narrow focus on "the language of science").

SAIL's approach to assessment reflects its approach to curriculum and instruction, described above. *SAIL* assessments center equity by scaffolding multilingual learners' engagement in cognitively demanding three-dimensional performances focused on making sense of phenomena. For example, in one assessment task anchored in the phenomenon of water quality in their city, students argue from evidence about whether there is a difference in quality between tap and bottled water. To scaffold their argument writing, students are provided a graphic organizer that includes response spaces for claim, evidence, and reasoning. In another task, students develop a model to explain the disappearance of a population from their local park ecosystem. To scaffold their engagement in modeling, students are reminded that their model-based explanation should address the "why" of the population's disappearance. These scaffolds are intended to ensure that English proficiency does not present a barrier to multilingual learners accessing the tasks in ways comparable to their peers, thus striving for "comparable levels of attainment" across student groups (NASEM, 2022, p. 23). Over the course of the yearlong curriculum, these scaffolds are removed as students become more proficient with science and engineering practices (e.g., arguing from evidence, modeling).

In addition, *SAIL* assessments expand what "counts" as evidence of science learning beyond traditionally privileged forms of expression. Assessments with multilingual learners have traditionally privileged written language and framed nonlinguistic modalities (e.g., visuals) as accommodations. In contrast, *SAIL* assessments elicit multimodal performances (e.g., visual, written, oral) as part of their design to capture a more complete and more nuanced picture of what students know and can do. For example, a student's visual representation of an ecosystem could offer insight into their science understanding in ways that would not otherwise be apparent from their written response alone (Grapin, 2022). Additionally, whereas assessments have traditionally privileged what students can do on their own (i.e., their independent performance), *SAIL* assessments foreground what students can do in dynamic interaction with others in their classroom community. For example, teachers ask probing questions (e.g., "What do you mean by…?") as they observe and interact with students working individually or in small groups. These questions are aimed at *hearing the science* in students' ideas (Grapin & Llosa, 2022), regardless of how those ideas are expressed (e.g., through gestures or everyday language). Then, teachers provide contingent feedback to advance students' science and language learning in tandem. In this way, *expansive approaches to assessment* are designed to capture multilingual learners' *expansive ways of making meaning* (Grapin, in press).

Currently, our research team is extending our work with *SAIL* to develop curriculum and assessments anchored in pressing societal challenges disproportionately impacting minoritized groups in society (Grapin et al., in press; Lee & Grapin, 2022). This emerging work centers justice in STEM learning with a focus on multilingual learners.

Further Reading

Advancing Equity and Strengthening Teaching with Elementary Mathematical Modeling: **https://www.eqstemm.org/**

Community for Advancing Discovery Research in Education (CADRE): Culturally Responsive STEM Education: **https://cadrek12.org/culturally-responsive-stem-education#advancing**

New York State Education Department webinar and brief series: **http://www.nysed.gov/bilingual-ed/integrating-science-and-language-all-students-focus-english-language-learners**

SAIL website: **https://www.nyusail.org**

Sol y agua Project: Developing a Culturally-Relevant Educational Game with a Theme of Water Sustainability: **https://emerging-researchers.org/projects/12542/**

Sol y aqua project: *Bilingual Culturally Relevant CS Curriculum* [video]. 2021 STEM For All Video Showcase **https://stemforall2021.videohall.com/presentations/2139**

References

Aguirre, J. M., Anhalt, C. O., Cortez, R., Turner, E. E., & Simic-Muller, K. (2019). Engaging teachers in the powerful combination of mathematical modeling and social justice: The Flint water task. *Mathematics Teacher Educator*, *7*(2), 7–26. **https://doi.org/10.5951/mathteaceduc.7.2.0007**

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. **https://doi.org/10.3102/0013189X07306523**

Bang, M. (2019). Making assessment responsive to culturally and linguistically diverse identities. Keynote Address at the 3rd National Council on Measurement in Education (NCME) Special Conference on Classroom Assessment, September 18–19, Boulder, CO.

Bang, M., Marin, A., & Medin, D. (2018). If Indigenous peoples stand with the sciences, will scientists stand with us? *Daedalus*, *147*(2), 148–159. **https://doi.org/10.1162/DAED_a_00498**

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*(6), 1008–1026. **https://doi.org/10.1002/sce.20392**

Bang, M., Warren, B., Rosebery, A. S., & Medin, D. (2013). Desettling expectations in science education. *Human Development*, *55*(5–6), 302–318. **https://doi.org/10.1159/000345322**

Basile, V., & López, E. (2015). And still I see no changes: Enduring views of students of color in science and mathematics education policy reports. *Science Education*, *99*(3), 519–548. **https://doi.org/10.1002/sce.21156**

Burgess, T., & Patterson Williams, A. (2022). Utilizing theory to elucidate the work of creating equity for transformation within the science classroom. *Science Education*, *106*(5), 1071–1083. **https://doi.org/10.1002/sce.21721**

Calabrese Barton, A. (1998). Reframing "Science for all" through the politics of poverty. *Educational Policy*, *12*(5), 525–541. **https://doi.org/10.1177/0895904898012005**

Calabrese Barton, A., Tan, E., & Birmingham, D. J. (2020). Rethinking high-leverage practices in justice-oriented ways. *Journal of Teacher Education*, *71*(4), 477–494. **https://doi.org/10.1177/0895904898012005004**

Campbell, T., & Lee, O. (2021). Instructional materials designed for *A framework for K–12 science education* and the next generation science standards: An introduction to the special issue. *Journal of Science Teacher Education*, *32*(7), 727–734. **https://doi.org/10.1080/1046560X.2021.1975359**

Delgado, R., & Stefancic, J. (2006). *Critical race theory: An introduction.* University Press.

Edelson, D. C., Reiser, B. J., McNeill, K. L., Mohan, A., Novak, M., Mohan, L., Affolter, R., McGill, T. A. W., Buck Bracey, Z. E., Deutch Noll, J., Kowalski, S. M., Novak, D., Lo, A. S., Landel, C., Krumm, A., Penuel, W. R., Van Horne, K., González-Howard, M., & Suárez, E. (2021). Developing research-based instructional materials to support large-scale transformation of science teaching and learning: The approach of the OpenSciEd Middle School Program. *Journal of Science Teacher Education*, *32*(7), 780–804. **https://doi.org/10.1080/1046560X.2021.1877457**

Fine, C. G. McC. (2022). Translanguaging interpretive power in formative assessment co-design: A catalyst for science teacher agentive shifts. *Journal of Language, Identity & Education*, *21*(3), 191–211. **https://doi.org/10.1080/15348458.2022.2058858**

Fine, C. G. M., & Furtak, E. M. (2020). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education*, *104*(3). **https://doi.org/10.1002/sce.21565**

Ford, M. J., & Forman, E. A. (2006). *Redefining Disciplinary Learning in Classroom Contexts. Review of Research in Education*, 30, 1–32. **http://www.jstor.org/stable/4129768**

Grapin, S. E. (2022). Are we missing part of the picture? Assessing English learners in science from a multimodal perspective. In M. A. Christison, J. Crandall, & D. Christian (Eds.), *Research on integrating language and content in diverse contexts* (pp. 38–52). Routledge. **https://www.routledge.com/Research-on-Integrating-Language-and-Content-in-Diverse-Contexts/Christison-Crandall-Christian/p/book/9781032013862**

Grapin, S. E. (in press). Assessment of English learners and their peers in the content areas: Expanding what "counts" as evidence of content learning. *Language Assessment Quarterly*, *2*, 215–234. **https://doi.org/10.1080/15434303.2022.2147072**

Grapin, S. E., Dudek, S., & Lee, O. (in press). Justice-centered STEM education with multilingual learners: Computational modeling to explain and design solutions to COVID-19 disparities. *Science Scope*.

Grapin, S. E., & Llosa, L. (2022). Multimodal tasks to assess English learners and their peers in science. *Educational Assessment, 27*(1), 46–70. **https://doi.org/10.1080/10627197.2022.2028139**

Gunckel, K. L. (2019). Repairing elementary school science. *Theory into Practice*, *58*(1), 71–79. **https://doi.org/10.1080/00405841.2018.1536918**

Gutiérrez, R. (2018). *The need to rehumanize mathematics.* In I. Goffney, R. Gutiérrez & M. Boston (Eds.), *Annual perspectives in mathematics education: Rehumanizing mathematics for Black, Indigenous, and Latinx students.* National Council of Teachers of Mathematics.

Guzman-Orth, D., Supalo, C. A., Smith, D. W., Lee, O., & King, T. (2021). Equitable STEM instruction and assessment: Accessibility and fairness considerations for special populations. *ETS Research Report Series*, *2021*(1), 1–16. **https://doi.org/10.1002/ets2.12324**

Harding, S. (2008). *Sciences from below: Feminisms, postcolonialities, and modernities*. Duke University Press.

Israel, M., Kester, B., Williams, J. J., & Ray, M. J. (2022). Equity and inclusion through UDL in K–6 computer science education: Perspectives of teachers and instructional coaches. *ACM Transactions on Computing Education, 22*(3), 1–22.

Kang, H. (2022). Teacher responsiveness that promotes equity in secondary science classrooms. *Cognition and Instruction*, *40*(2), 206–232. **https://doi.org/10.1080/07370008.2021.1972423**

Kang, H., & Furtak, E. M. (2021). Learning theory, classroom assessment, and equity. *Educational Measurement: Issues and Practice*, *40*(3), 73–82. **https://doi.org/10.1111/emip.12423**

Kang, H., Thompson, J., & Windschuttle, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, *98*(4), 674–704. **https://doi.org/10.1002/sce.21123**

Kang, H., Talafian, H., & Tschida, P. (2022). Expanding opportunities to learn in secondary science classrooms using unconventional forms of classroom assessments. *Journal of Research in Science Teaching*. **https://doi.org/10.1002/tea.21824**

Kimmerer, R. W. (2013). *Braiding sweetgrass*. Milkweed Editions. K–12 Computer Science Framework. (2016). **http://www.k12cs.org.**

Learning in Places Collaborative. (2020). *Learning in places*. **http://learninginplaces.org/**

Leckey, E. H., Littrell, M. K., Okochi, C., González-Bascó, I., Gold, A., & Rosales-Collins, S. (2021). Exploring local environmental change through filmmaking: The Lentes en Cambio Climático program. *The Journal of Environmental Education*, *52*(4), 207–222. **https://doi.org/10.1080/00958964.2021.1949570**

Lee, C. D. (2017). An ecological framework for enacting culturally sustaining pedagogy. In D. Paris & H. S. Alim (Eds.), *Culturally sustaining pedagogies: Teaching and learning for justice in a changing world* (pp. 261–273). Teachers College Press.

Lee, O., Llosa, L., Grapin, S., Haas, A., & Goggins, M. (2019). Science and language integration with English learners: A conceptual framework guiding instructional materials development. *Science Education*, *103*(2), 317–337. **https://doi.org/10.1002/sce.21498**

Lee, O., & Grapin, S. E. (2022). The role of phenomena and problems in science and STEM education: Traditional, contemporary, and future approaches. *Journal of Research in Science Teaching, 59*(7), 1301–1309. **https://doi.org/10.1002/tea.21776**

Lee, O., Llosa, L., Grapin, S. E., Haas, A., & Goggins, M. (2019). Science and language integration with English learners: A conceptual framework guiding instructional materials development. *Science Education, 103*(2), 317–337. **https://doi.org/10.1002/sce.21498**

Lee, O., Quinn, H., & Valdes, G. (2013). Science and language for English language learners in relation to next generation science standards and with implications for common core state standards for English language arts and mathematics. *Educational Researcher*, *42*(4), 223–233. **https://doi.org/10.3102/0013189X13480524**

Likely, R. (2020). *"Lotions and potions": Exploring Black girls' engagement in and perceptions of science practices through hair care* [Unpublished doctoral dissertation]. Drexel University, School of Education.

Llosa, L., Grapin, S. E., & Haas, A. (2022). Four types of embedded formative assessment in the science classroom with English learners. *Science and Children, 59*(3), 58–63. **https://www.nsta.org/science-and-children/science-and-children-januaryfebruary-2022/assessment-english-learners**

Lo, A. S., Glidewell, L., O'Connor, K., Allen, Anna-Ruth, Herrmann-Abell, C. F., Penuel, W. R., Wingert, K., & Lindsay, W. (2022). Promoting shifts in teachers' understanding and use of phenomena in instruction and assessment. In C. Chinn, E. Tan, C. Chan, & Y. Kali. *(Eds.). Proceedings of the 16th International Conference of the Learning Sciences – ICLS 2022* (pp. 1145–1148)*.* International Society of the Learning Sciences.

Love, B. L. (2019). *We want to do more than survive*. Beacon.

McKinney de Royston, M., & Farinde, A. A. (2018). Race and poverty in science education: Questions and tensions for the field. *Counterpoints*, *442*, 193–214. **https://www.jstor.org/stable/45212201**

Mehta, J., & Fine, S. (2019). *In search of deeper learning: The quest to remake the American high school.* Harvard University Press. **https://doi.org/10.4159/9780674239951**

Morales-Doyle, D. (2017). Justice-centered science pedagogy: A catalyst for academic achievement and social transformation. *Science Education*, *101*(6), 1034–1060. **https://doi.org/10.1002/sce.21305**

Nasir, N., McKinney de Royston, M., Barron, B., Bell, P., Pea, R., Stevens, R., & Goldman, S., (2020). Learning pathways: How learning is culturally organized. In N. Nasir, C. Lee, R. Pea & M. McKinney de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 195–211). Routledge.

National Academies of Sciences, Engineering, and Medicine. (2018a). *English learners in STEM subjects. Transforming classrooms, schools, and lives*. National Academies Press.

National Academies of Science, Engineering, and Medicine. (2018b). *How people learn II: Learners, contexts and cultures*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2022). *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators*. National Academies Press.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics.*

National Research Council. (2012). *A framework for K–12 science education.* National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards.* The National Academies Press.

National Research Council. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

National Science Foundation. (2020). Discovery Research PreK-12 (DRK-12): Program Solicitation: NSF 20-572. (**https://www.nsf.gov/pubs/2020/nsf20572/nsf20572.htm**).

NGSS Lead States. (2013a). *Next generation science standards: For states, by states.* National Academies Press.

NGSS Lead States. (2013b). *Next generation science standards: For states, by states*. *Appendix D–All standards, all students: Making the NGSS accessible to all students.* **https://www.nextgenscience.org/resources/ngss-appendices**

Noble, T., Suarez, C., Rosebery, A., Connor, M. C. O., Warren, B., & Hudicourt-barnes, J. (2012). *"I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science*. *49*(6), 778–803. **https://doi.org/10.1002/tea.21026**

Penuel, W. R., Horne, K. V., Turner, M. L., Jacobs, J. K., & Sumner, T. (2019). Developing tasks to assess phenomenon-based science learning: Challenges and lessons learned from building proximal transfer tasks. *Science Education*, *103*(6), 1367–1395. **https://doi.org/10.1002/sce.21544**

Philip, T. M., & Azevedo, F. S. (2017). Everyday science learning and equity: Mapping the contested terrain. *Science Education*, *101*(4), 526–532. **https://doi.org/10.1002/sce.21286**

Randall, J. (2021). Color neutral is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82–90. **https://doi.org/10.1111/emip.12429**

Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and informative for all: Universal design for learning (UDL) and the future of assessment. *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (pp. 167–180). Springer. DOI:10.1007/978-3-319-71126-3_11

Semken, S., Ward, E. G., Moosavi, S., & Chinn, P. W. U. (2017). Place-based education in geoscience: Theory, research, practice, and assessment. *Journal of Geoscience Education*, *65*(4), 542–562. **https://doi.org/10.5408/17-276.1**

Shepard, L. A. (2021, Fall). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *The American Educator, 45*(3). **https://www.aft.org/ae/fall2021/shepard**

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34. **https://doi.org/10.1111/emip.12189**

Taylor, C. S. (2022). *Culturally and socially responsible assessment: Theory, research, and practice*. Teachers College Press.

Tzou, C., Bang, M., & Bricker, L. (2021). Commentary: Designing science instructional materials that contribute to more just, equitable, and culturally thriving learning and teaching in science education. *Journal of Science Teacher Education*, *32*(7), 858–864. **https://doi.org/10.1080/1046560X.2021.1964786**

Tzou, C., Meixi, Suárez, E., Bell, P., LaBonte, D., Starks, E., & Bang, M. (2019). Storywork in STEM-Art: Making, materiality and robotics within everyday acts of Indigenous presence and resurgence. *Cognition and Instruction*, *37*(3), 306–326. **https://doi.org/10.1080/07370008.2019.1624547**

Warren, B., Vossoughi, S., Rosebery, A. S., Bang, M., & Taylor, E. V. (2020). Multiple ways of knowing: Re-imagining disciplinary learning. In N. S. Nasir, C. D. Lee, R. D. Pea, & McKinney de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 277–294). Routledge.

Wilson-Lopez, A., Mejia, J. A., Hasbún, I. M., & Kasun, G. S. (2016). Latina/o adolescents' funds of knowledge related to engineering. *Journal of Engineering Education*, *105*(2), 278–311. **https://doi.org/10.1002/jee.20117**

Wilson-Lopez, A., Sias, C., Smithee, A., & Hasbún, I. M. (2018). Forms of science capital mobilized in adolescents' engineering projects. *Journal of Research in Science Teaching*, *55*(2), 246–270. **https://doi.org/10.1002/tea.21418**

Wright, C., & Riley, A. (2021). Mitigating the need for resiliency for Black girls: Reimagining the cultural brokering through a lens of science as white property. *Cultural Studies of Science Education*, *16*(2), 495–500. **https://doi.org/10.1007/s11422-020-10005-9**

Suggested Citation

Furtak, E. M., & Lee, O. (2023). Equity and justice in classroom assessment of STEM learning. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives* (pp. 83–94). Community for Advancing Discovery Research in Education (CADRE). Education Development Center.

Teacher Knowledge and Practices for Assessment

Ryan Seth Jones, Middle Tennessee State University,
Eric Banilower, Horizon Research, Inc., and
Shuchi Grover, Looking Glass Ventures

4

Introduction

Teaching is a practice with heavy epistemic demands. A teacher is often in the same position as a physician where the efficacy of the practitioner's work heavily depends on a diagnosis of invisible (or difficult to see) characteristics of a person. Teachers' assessments of their students' thinking are always assessments of traits they cannot see, which requires a constant coordination of the visible with the invisible (Pellegrino et al., 2001). Since a teacher's job is to support students to develop competencies, teachers regularly need to use various forms of assessment to generate, interpret, and use evidence about students' thinking, abilities, and dispositions. High-quality assessment is necessary for responsive instruction, just as diagnosis is necessary for a physician's care to be responsive to patients, which is why this type of assessment is often called assessment *for* learning, in contrast to assessment *of* learning (Black & Wiliam, 2010).

Assessment for learning, often referred to as *formative assessment* or *classroom-based assessment*, is a fundamental practice for teachers with significant impact on students' learning (Black & Wiliam, 1998; National Research Council, 2001; Kingston & Nash, 2011; Furtak et al., 2016). This is likely why 67 percent of eighth-grade mathematics teachers reported that they adjust their instruction multiple times a week in response to students' learning, and around half of all high school math and science teachers reported that formative assessment was a significant focus of their professional development in recent years (NAEP, 2021; Banilower et al, 2018).

A significant proportion of teachers' work involved the design and use of assessment strategies and the interpretation and use of the evidence generated by these strategies. Although talk of "assessment" can often conjure images of "tests" or "quizzes," high-quality formative assessment involves a diverse set of strategies deployed almost continuously over the course of instruction, such as questions during discussions and tasks that elicit evidence of students' thinking. This makes assessment not only fundamental to teaching because of its importance, but also because of the nearly continuous need for assessment to guide instruction.

We use the word *practice* here intentionally to emphasize that the nature of assessment work is a coordinated set of actions, ideas, interpretations, and meanings that is continually built and maintained by a community with shared goals in mind (Furtak et al., 2019; Furtak at al., 2016; Grossman et al., 2009; Lampert, 2010). Like many practices, assessment practice is a relational practice in which teachers aim to coordinate their work with students in their care and professional and disciplinary standards for their practice (Gess-Newsome, 2015; Lampert, 2010). This means that assessment practice can never be reduced to a set of abstract skills or ideas but is always maintained in relation to the people, tools, and goals in particular communities (Grossman et al., 2009). Teachers' assessment practice is always situated within local, professional, and political networks that both support and constrain what is possible. This means researchers, policymakers, and practitioners all must work to support teachers to engage in high-quality assessment practices.

In this section, we argue that teachers need support in three domains in order to carry out high-quality assessment in their classrooms. Before describing these domains, we elaborate on the work of classroom assessment and what teachers need in order to be able to do this work. As we have stated above, we believe it is impossible to describe this work as an abstract set of skills, so we do not mean to suggest that this work is invariant across contexts. Rather, we will attempt to articulate generalized principles of the practice that are always contingent on the ground and in relationship with the people, tools, and values of the social networks that teachers participate in. After this elaboration, we will argue for three types of support teachers need to carry out this work. We view these forms of support as the tools teachers need from the professional teaching community to support the challenging practice of assessment (Furtak et al., 2019). In addition, we illustrate our argument using an example of a research program that supports curriculum and assessment for middle school data and statistics.

Teachers' Work in the Practice of Assessment

Teachers work to develop students' thinking, which means the primary target of the teachers' work occurs on mechanisms that are beyond a teachers' ability to observe directly. Although learning goals sometimes include words and inscriptions to remember, the goals extend far beyond memorization. In STEM education, goals for students' thinking are often characterized as concepts for students to understand. This means that students do not merely remember the symbols that refer to the concept (e.g. words, images, or other inscriptions), but that they ascribe meaning to those symbols that have meaningful correspondence to both disciplinary meanings and the students' worlds and experiences (Carlsen, 2007). Over the last few decades, STEM education communities have developed an additional type of learning goal, the development of

STEM practices. Although practices might seem more visible since they are always carried out interactionally with others in a community, assessments about student competencies in a practice require information about a student beyond just their actions in one context. This means that competency in practices, like concepts, are invisible traits of students that teachers need to make inferences about.

Because of this, teachers' assessment practice first and foremost involves *designing and enacting strategies to manifest invisible student characteristics in observable actions* (Furtak et al., 2019; Grover, 2021; Heritage & Wylie, 2019). This work is often referred to as *eliciting* [evidence of] student thinking (e.g. Franke & Kazemi, 2001). Strategies to elicit student thinking can take on many forms. Tests and quizzes are so tightly associated with most conceptions of assessment because they are two strategies to elicit student thinking that often dominate teachers' assessment practices. Other strategies that are often better suited to solicit more detailed information about students' thinking include individual and group tasks that ask students to respond to more open-ended or complex scenarios and class discussions about phenomena where targeted learning goals might be used by students to make sense of it. These strategies can create a space in which students not only share (make visible) their thinking with the teacher, but also with other students in the class.

Formative assessment strategies to elicit thinking are best used during moments of instruction where students are in the process of learning. Students' ideas often shift and change as they see and hear other students' strategies and ideas, providing an opportunity for a teacher to observe substantive and detailed evidence of changes in students' thinking as students work on the task and the conversation progresses (Heritage & Wylie, 2019; Kang et al., 2014). This scenario highlights the dynamic, community nature of classroom-based assessment whereby elicitation of student thinking provides both evidence for the teacher but also learning opportunities for all students in the classroom.

When the assessment target is a STEM practice, the teacher must design scenarios that engage students in the practice in a meaningful and accessible way to assess proficiency in the practice. Since practices are made up of actions and ideas that are carried out interactionally in communities with shared goals, these assessment scenarios are often designed to engage students in approximations to the types of interactions STEM professionals engage in to generate and revise knowledge (Ford, 2015). For example, students might present to the class their findings from an investigation as the teacher facilitates a conversation meant to approximate the practice of scientific argumentation and critique. This means the target of the assessment might not be the concepts that students are using in their conversation, but might be focused on the epistemic practice of how the students justify their reasoning, support claims with evidence, or revise claims in light of legitimate critique from others.

These broader elicitations of student thinking must be matched with more focused probes in order for teachers to make high-quality assessments of students' learning. Teachers need to be able to *design and ask questions that elicit responses* related to specific ideas, allow students opportunities to elaborate on their own thinking, or ask students to consider other students' ideas and compare them to their own (Black et al., 2011). For these types of interactions to occur, teachers must ask more than evaluative questions that often ask students to guess what the teacher is thinking (Alonzo, 2018; Gotwals & Birmingham, 2016). They must ask open-ended questions that provide opportunities for students to consider ideas from multiple perspectives, make use of ideas in new and sometimes unfamiliar contexts, draw on personal and cultural ideas, and engage in extended conversation about the questions (Black et al., 2011). In fact, well-designed questions paired with classroom norms that support these types of conversations can often lead to discussions in which students begin asking one another questions that further both assessment and learning goals.

The work a teacher must do to design assessment scenarios to elicit student thinking often happens during planning and before classroom instruction. However, the work of questioning happens both in planning and during instruction. To develop good questions during planning, teachers must anticipate the types of responses students are likely to give and how the questions might leverage these responses to further elicit student thinking (Stein et al., 2008). However, questioning work during classroom instruction always has an improvisational nature to it as teachers respond in the moment to dynamic student strategies and conversations, sometimes referred to as *orchestration* (Looi & Song, 2013). Orchestration requires more than just planning well—it requires the ability to closely observe student actions, listen carefully to student comments, and respond to them with the types of questions that will continue to elicit high-quality evidence of their thinking and competencies.

This means the third practice teachers need to engage in is *interpreting observable actions to make valid inferences about invisible student traits*. Assessment validity can often refer to the types of evidence that psychometricians collect to justify inferences made with large-scale assessments, but in formative assessment the teacher must determine what counts as valid evidence for claims about student thinking (Pellegrino et al., 2016). Of course, validity evidence of tasks can be generated by researchers and curriculum developers, which we will discuss in the following section, but the teacher always plays an important role on the ground in generating valid interpretations of assessment evidence and using assessment tools in ways that generate valid inferences about their students. This requires a deep knowledge of the target discipline, the multiple ways students think about these ideas, and the types of observable actions, such as words and inscriptions, that serve as valid evidence of particular types of thinking.

Because of the challenging nature of these assessment practices, they are best carried out iteratively and in collaboration with other teachers and professional communities (Grover, 2021). Task design, questioning, and interpretation can all be iteratively improved as teachers reflect on their assessment efforts and revise their strategies for the future. For example, maybe the first time a teacher uses an assessment strategy there is very little known about likely student responses, and only upon reflection does the teacher notice some important ideas that they were blind to during instruction. This reflection can be used to prepare to better recognize and use these ideas in future assessment efforts. Alternatively, collaborative groups of teachers working together to plan for assessments or to interpret students' responses to assessments can improve the quality of the assessment work (Norwich et al., 2014). Multiple perspectives can help teachers to notice more detail about their students than they would have on their own. And last, the resources created by researchers and curriculum developers in teachers' professional communities can serve a crucial role in supporting teachers to engage in high-quality assessment practices.

Necessary Support for Teachers' Assessment Practice

Teachers carry out formative assessment practices on the ground during their ongoing instruction. However, these practices do not exist in isolation, but within networks of colleagues, professional organizations, and research communities. These communities, then, have the responsibility to generate resources that support the hard work of high-quality formative assessment practices.

To begin, teachers need *disciplinary-specific models of student cognition that are empirically grounded and developmental to carry out high-quality assessment* (Black et al., 2011; Pellegrino et al., 2016). Of course, all assessment inferences will be guided by a model of student thinking, even if just implicit. Many times, when left implicit, the model of student thinking teachers use can be a simplistic framework of "correct" or "incorrect" ideas (Alonzo, 2018; Gotwals & Birmingham, 2016). This is why teachers need models of student thinking and learning that are supported by empirical evidence about students' sense making. These models can serve to support teachers, curriculum developers, assessment developers, and researchers to share a vision of student learning that is grounded in research in the particular domain. This shared vision, then, can serve as a boundary object (Star & Griesemer, 1989) that curriculum, assessment, and teachers' practice can be organized around to support evidence-based assessment practices (Black et al., 2011; Pellegrino et al., 2016). Of course, these models are always hypothetical because student thinking can't be predicted with certainty, and there will always be variation from context to context in students' responses to particular instructional strategies (Simon, 1995). A model of student cognition will always be a potential route to learning (Clements & Sarama, 2004). However, since they are grounded in empirical evidence of students' actual thinking while learning, they describe the ways of thinking that teachers can expect to see during instruction.

Models of student cognition are developmental and thus can provide a road map that not only articulates the end learning goal, which is often represented in standards and objectives, but also the learning trajectory by which students might travel to arrive at that goal (Black et al., 2011; Grover, 2021; Pellegrino et al., 2016). A developmental trajectory that is empirically grounded in student sense making, developmental, and disciplinary specific would articulate important ideas that students have been shown to make use of when learning a specific idea or competency. This means it would be informed by disciplinary-specific ideas and learning goals. More background on learning trajectories (also referred to as learning progressions) can be found in **Section 1. Connecting Classroom Assessment with Learning Goals and Instruction through Theories of Learning** in this report.

Students' trajectories toward understanding these should also be articulated in the language and ideas students use as they make sense of a question or phenomenon. After all, students' sense making often involves ideas and strategies that disciplines don't recognize because of the culturally constrained ways that teaching and learning theory has historically developed. Thus, students often grapple with disciplinary knowledge in ways that illuminate key nuances in understanding obscured by disciplinary convention (Lehrer, et al., 2014). If only conventional disciplinary language, representations, or ways of thinking are used to develop these models then researchers will likely ignore the powerful and transformative ways students think (Lehrer & Schauble, 2015). This consideration is important because, as **Section 3. Equity and Justice in Classroom Assessment of STEM Learning** shows, current formative assessment practices often don't adequately serve minoritized populations because they often ignore unique cultural resources and practices these students bring to the classroom. Instead, these models should support teachers to develop a vision for a wide range of student ideas, how they are productive for making sense of the question at hand, and how they are related to disciplinary ways of thinking.

In order to make judgments about students' thinking in relation to these models, teachers also need *feasible strategies and tools for eliciting observable evidence of student thinking* (Furtak et al., 2019; Heritage & Wylie, 2019). This requires that assessment tools be integrated with the overall curriculum design so that teachers can coordinate the evidence generated by the tools with specific instructional strategies. Although teachers need to be competent in designing tools themselves, research-based tasks and strategies are an important resource for teachers to draw upon in their work.

These tools can often take the form of prompts, investigations, or tasks that organize classroom activity in ways that elicit diverse student ideas and strategies (Furtak et al., 2019). Tasks engage students with a scenario that the designer believes will elicit student thinking about a phenomenon or concept and generate responses to the scenario that inform teachers' assessments. In addition, tasks can engage students in STEM practices such as argumentation. One advantage of tasks that are developed by curriculum designers or researchers is that they can be tested and revised in response to empirical evidence. This has the potential to improve the validity of the task to elicit relevant evidence because they have been revised in response to how they performed when used with students.

Tools to elicit student thinking can also take the form of teaching strategies to be deployed in the course of instruction. For example, questions can be used to encourage students to further elaborate their thinking during classroom discussions. Although questions are often deployed in the moment, which requires significant skill and improvisation, they can also be developed before instruction in anticipation of the conversation. Questions then are not only extemporaneous, but they are strategies teachers can prepare for using question developed by curriculum developers and researchers.

Lastly, teachers need *tools that support interpretations of the generated evidence* in terms of the cognitive model (Confrey & Shah, 2021). Even with high-quality assessment tools grounded in empirical and developmental cognitive models, the student responses elicited by the assessment strategy will always be highly variable with a fair dose of idiosyncrasy. This means coherent connections among assessment responses, the cognitive model, and instructional implications are always a challenge (Black et al., 2011). Teachers need support to make valid inferences and to use these inferences to inform their teaching.

These tools are often supported or instantiated in technology. These technology tools can provide an interactive context for a student-driven task, elicitation of their thinking, or producing interactive probes that further uncover student understanding. They can also be used independent of direct teacher involvement or in concert with real-time instruction. However, teachers are always at the forefront of the decision-making of why, when, and where these tools are deployed. Technology can also be deployed to help harvest, display, and interpret results of student work. More on the use of technology for assessment can be found in **Section 5. Technology-Based Innovative Assessment** of this report.

This type of support requires clear and coherent coordination among the cognitive model, curriculum, and assessment tools. In addition, teachers need support to develop an understanding of these connections. A learning trajectory approach to research and development has the potential to create such a system of support. Although learning trajectories and learning progressions have been conceptualized and studied in many different ways, most approaches conceptualize a model of student learning and the conditions under which it occurs as interrelated and design assessment tools to inform teachers about student thinking in the language of the model of cognition. Although the model is sometimes referred to as the *learning trajectory*, it is in fact the arrangement among model, curriculum, and assessment that constitutes the learning trajectory (Lehrer et al., 2014). Previous sections in this report elaborate more on learning trajectories, but it is worth noting here that this approach provides much needed support to teachers for high-quality assessment practices.

Box 4: Example from the DRK-12 Portfolio: Illustration: A Learning Progression to Support the Development of Statistical Reasoning

Lehrer et al. (2014) describe a learning progression approach to supporting teachers' formative assessment practices related to statistical reasoning, and they articulate a vision for a program of research that coordinates models of student learning, assessment tools, and curricular implications. In their view:

A progression must be articulated in ways that support the alignment of discipline, learning, instruction, and assessment. An epistemic view of discipline considers how concepts are generated and refined in a field of inquiry. Depictions of learning include descriptions of states of student knowledge, including concepts and practices, and consequential transitions among these states. If learning progressions are to guide teaching, then teachers must be able to identify classes of student performances as representing particular states of the progression. This effort may require innovations in assessment that reveal students' ways of thinking. And, once student states are identified, teaching practices must come to include a repertoire of appropriate pedagogical responses to support students' progress. The teacher's vision of the disciplinary material to be taught might also need to shift from a logical decomposition of subject matter to a focus on how learners are disposed toward the creation and revision of that subject matter. (p. 33-34)

Their system coordinates learning, instruction, and assessment to support the development of statistical reasoning in late elementary and middle grades students and is organized around six dimensions that serve as the model of student cognition. Lehrer and Schauble (2015) refer to their representations of these dimensions as *construct maps*. The details of these six construct maps are beyond the scope of this section (and readers are encouraged to explore them further in the original source (**https://www.researchgate.net/publication/275156599_Learning_Progressions_The_Whole_World_is_NOT_a_Stage**), but we will highlight one construct map, data display, and how this system supports teachers' assessment practices.

The curricular design regularly asks students to invent novel approaches to data visualization, statistics, and modeling in order to elicit their thinking about data visualization. When students create these approaches, they often deviate from conventions described in standards (e.g., histograms, dot plots) but provide information to teachers so they can make assessments about the ideas students are making use of when they create the data visualizations. For example, does a student think that values that fall within the range of our data, but not measured by anyone in our sample, be included in our scale? The cognitive model articulated in the data display construct map describes common ways of thinking in an effort to support teachers to assess and respond to these student created data displays. It provides teachers a trajectory of possible ways of thinking and supports teachers in looking for fundamental display principles, such as order, frequency, and scale, in students' strategies to visually display their data. As Lehrer and Schauble (2015) describe, in this system "teachers do not leave the comparison of displays to happenstance, but instead promote particular classes of comparison, such as comparing a display that partitions continuous data but does not reveal its scale to one that partitions and preserves scale." (p. 37-38)

Wilson and Lehrer collaborated to use Wilson's construct-centered assessment design (Wilson, 2005) to develop formative assessment tasks to inform teachers about students' evolving thinking as they engage with the curricular design. These assessment items included multiple-choice and free response items and were developed across iterative cycles of item writing, field testing, scoring, and psychometric modeling. Through this process, the items were continually revised, and at times the construct map was revised to represent new types of student thinking the items elicited. These items were then integrated into curriculum materials and professional development to support teachers to elicit relevant ways of thinking, interpret student products in terms of the construct maps, and then make instructional decisions based on those interpretations. This type of assessment work has been extended into other content areas in the NSF DRK-12 collaborative project (NSF 1621265, 0733334).

This research program has motivated and informed ongoing design and research. For example, Confrey (NSF 1621254) led the development of a digital learning map

that drew on Lehrer's construct maps and curricular design, but created innovative digital assessment tools and representations of the construct maps.
In addition, Confrey's team further elaborated on the data display construct map to develop additional dimensions of student learning about data displays (Arnold et al., 2018; Confrey & Shah, 2021).

Illustration: Supporting Teachers in Implementing Research-Based Formative Assessment Practices

Sevian and colleagues have been developing and studying a professional development model for improving chemistry teachers' formative assessment practices in their NSF-supported DRK-12 project (NSF 1621228). This work also builds on a learning progression framework for chemical thinking (Sevian & Talanquer, 2014). The model of student cognition is organized around constructs (what they refer to as progress variables) that describe increasingly sophisticated ideas about key questions that chemistry professionals tend to ask (i.e. how do properties of matter emerge?).
They define increasingly sophisticated ideas in terms of the assumptions that students make about the substance under investigation and different modes of reasoning students make use of (i.e. relational versus descriptive).

This team also studied teachers' formative assessment enactment when using their assessment tools (Dini et al., 2020). They found that teachers' practices either take an evaluative stance where they judge student thinking in reference to alignment (or not) with conventional models or a more interpretive stance where they seek to understand the sense that the explanation makes to the student. The evaluative stance often closed down the conversation and limited assessment opportunities, while the interpretive stance opened up the conversation to elicit more details about students' ideas and provided a chance for teachers to advance student thinking.

The project has also pulled together several formative assessment tools related to the various elements of the chemical thinking learning progression and made these tools and its professional development program available to teachers through its website (**https://www.chemedx.org/ACCT**), providing a model for supporting teachers in using formative assessment within STEM education.

Summary and Future Directions

This section describes a vision for classroom-based formative assessment embedded in everyday but powerful teaching routines. Further, it lays out the work that is necessary for formative assessment: creating high-quality embedded assessment tasks, implementing them, and interpreting results in ways that are useful for guiding instruction. It also highlights several areas of research needed to support future efforts in this area, including the development of models of cognition for many topics addressed in the K–12 STEM curriculum, classroom-based assessment tools based on these models, and supports for teachers to coordinate their interpretations of assessment evidence in relation to the cognitive model in ways that inform instruction. The first and second are described previously in this section, including examples of projects within the DRK-12 portfolio working in these areas. The remainder of this section focuses on some of the challenges associated with a large-scale approach to the third area: teacher support.

There are a number of challenges in regard to supporting teachers in implementing high-quality classroom-based assessment. One is the sheer scale of the K–12 education system–based on data from the 2018 NSSME (Banilower et al., 2018), there are over 1.2 million science teachers and a somewhat greater number of mathematics teachers in the United States. Another issue is the relatively constant influx of new teachers into the profession, regardless of whether it is due to retirements or the "churn" of teachers leaving the profession, particularly in schools and districts that place greater demands on new teachers (Ingersoll et al., 2018). Teacher preparation for new disciplines or subjects, such as CS in K–12, involve unique challenges since most teachers themselves have no prior experience with computing and programming. As the focus moves from curriculum design and pedagogy to classroom teaching practice, attention to teacher professional development, professional communities, as well as resources focused on formative assessment (in addition to curriculum and pedagogy) are crucial.

Further, the STEM teaching force is not monolithic in terms of their preparation for teaching their subject and professional learning needs. For example, about three-quarters of science and mathematics teachers work at the elementary grades level, teaching in self-contained classrooms (i.e., they are responsible for teaching all core subjects) and are typically generalists without extensive training in specific subject areas (Banilower et al., 2018). Even at the middle and high school levels, where most teachers have substantive training in their discipline (i.e., science, mathematics), they may not have the deep knowledge of all of the content areas they are expected to teach. For example, many middle school science courses addressed multiple areas of science, such as earth science, life science, and physical science, and many high school science teachers are, due to staffing necessities, assigned one or more courses out of their primary field of expertise (e.g., a biology teacher assigned a chemistry course). This phenomenon is very common in CS and engineering as only small percentages of teachers have substantial training in these disciplines or how to teach them.

Thus, for this vision of classroom-based assessment to be achieved, it may be helpful in future research to articulate a theory of change that takes a broad view of the education system and identifies high-leverage strategies that can bring about change in classroom-based assessment practices. For example, previous NSF-funded work focused on the adoption of high-quality instructional materials and supporting teacher implementation through professional development focused on those materials as a scalable strategy (Weiss & Pasley, 2006). This approach, enhanced by more recent work around educative curriculum materials as a means to support teachers and their professional learning (e.g., see Davis et al., 2017) has the potential to facilitate the uptake of the research into learning progressions and models of student cognition more broadly.

Although the challenges with supporting teacher formative assessment on a large scale are significant, the research highlighted in this section has provided strong evidence that supporting teachers to engage in high-quality formative assessment practices is a worthwhile focus for improving STEM teaching and learning.

Major Implications:

For NSF and researchers:

- Assessment projects should attend to the knowledge and practices teachers need support in and design assessment systems to support these practices.
- Cognitive models should be explicitly articulated in ways that they can be examined apart from the assessment tools designed to elicit thinking, and so that teachers can understand them and make use of them.

Practitioners:

- Classroom-based assessment should be focused on understanding students' sensemaking and ideas rather than on only judging students' ideas in relation to conventional ideas and models that professional communities use.
- Eliciting student ideas should produce a variety of ways of thinking, so they should be designed to allow for diversity of strategies and approaches.
- Interpreting and making use of student ideas takes significant effort and is often best done in collaboration with other teachers and iteratively across time.

References

Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education*, *31*(2), 104–112.

Arnold, P., Confrey, J., Jones, R. S., Lee, H. S., & Pfannkuch, M. (2018). Statistics learning trajectories. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research statistics in education* (pp. 295–326). Springer International Publishing. **https://doi.org/10.1007/978-3-319-66195-7_9**

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). *Report of the 2018 NSSME+.* Horizon Research, Inc.

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. Assessment in Education. Assessment in Education, 5, 7-74. **http://dx.doi.org/10.1080/0969595980050102**

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspective, 9*(2–3), 71–123.

Carlsen, W. S. (2007). Language and science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 57–74). Erlbaum.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*, 81–89. doi:10.1207/s15327833mtl0602_1

Confrey, J., & Shah, M. (2021). Examining instructional change at scale using data from diagnostic assessments built on learning trajectories. *ZDM–Mathematics Education*, *53*(6), 1265–1283.

Davis, E. A., Palincsar, A. S., Smith, P. S., Arias, A. M., & Kademian, S. M. (2017). Educative curriculum materials: Uptake, impact, and implications for research and design. *Educational Researcher, 46*(6), 293–304. **https://doi.org/10.3102/0013189X17727502**

Dini, V., Sevian, H., Caushi, K., & Orduña Picón, R. (2020). Characterizing the formative assessment enactment of experienced science teachers. *Science Education*, *104*(2), 290–325.

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the next generation science standards. *Science Education*, *99*(6), 1041–1048.

Franke, M. L., & Kazemi, E. (2001). Learning to teach mathematics: Focus on student thinking. *Theory into Practice*, *40*(2), 102–109.

Furtak, E. M., Heredia, S. C., & Morrison, D. (2019). Formative assessment in science education: Mapping a shifting terrain. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 97–125). Routledge.

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science*, *44*(3), 267–291.

Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK. In A. Berry, P. J. Friedrichsen, & J. Loughran (Eds.), *Teaching and learning in science series. Reexamining pedagogical content knowledge in science education*. (pp. 28–41). Routledge.

Gotwals, A. W., & Birmingham, D. (2016). Eliciting, identifying, interpreting, and responding to students' ideas: Teacher candidates' growth in formative assessment practices. *Research in Science Education*, *46*, 365–388. DOI:10.1007/s11165-015-9461-2

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers college record*, *111*(9), 2055–2100.

Grover, S. (2021). Toward a framework for formative assessment of conceptual learning in K–12 computer science classrooms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 31–37).

Heritage, M., & Wylie, C. (2019). Teacher preparation in mathematics. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines.* (pp. 207–242). Routledge.

Ingersoll, R., Merrill, E., Stuckey, D., & Collins, G. (2018). Seven trends: The transformation of the teaching Force (Research report (#RR 2018–2). Consortium for Policy Research in Education, University of Pennsylvania.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education, 98*(4), 674–704.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*, 28–37. **https://doi.org/10.1111/j.1745-3992.2011.00220.x**

Lampert, M. (2010). Learning teaching in, from, and for practice: What do we mean? *Journal of Teacher Education*, *61*(1–2), 21–34.

Lehrer, R., Kim, M. J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. P. Maloney, J. Confrey, & K. H. Nyguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (31–60). Information Age Publishing.

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education*, *99*(3), 432–437. **https://www.researchgate.net/publication/275156599_Learning_Progressions_The_Whole_World_is_NOT_a_Stage**

Looi, C.-K., & Song, Y. (2013). Orchestration in a networked classroom: Where the teacher's real-time enactment matters. *Computers & Education, 69*, 510–513. doi: **http://dx.doi.org/10.1016/j.compedu.2013.04.005**

National Assessment of Educational Progress. (2021). 2021 Mathematics Assessment. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

National Research Council. (2001). Adding it up: Helping children learn mathematics. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). National Academy Press. **https://doi.org/10.17226/9822.**

Norwich, B., Dudley, P., & Ylonen, A. (2014). Using lesson study to assess pupils' learning difficulties. *International Journal for Lesson and Learning Studies, 3*(2).

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.

Sevian, H. & Talanquer, T. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice, 15*(1), 10–23.

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145.

Star, S. L. & Griesemer, J. (1989). Institutional ecology, 'translations', and boundary objects: Amateurs and professionals on Berkeley's museum of vertebrate zoology. *Social Studies of Science 19*(3), 387–420.

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.

Weiss, I. R. & Pasley, J. D. (2006, March). Scaling up instructional improvement through teacher professional development: Insights from the local systemic change initiative. *CPRE Policy Briefs, RB-44*.

Suggested Citation

Jones, R. S., Banilower, E., & Grover, S. (2023). Teacher knowledge and practices for assessment. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives* (pp. 83–94). Community for Advancing Discovery Research in Education, Education Development Center, Inc.

Technology-Based Innovative Assessment

Xiaoming Zhai, University of Georgia, and
Eric Wiebe, North Carolina State University

5

This section presents an overview of critical developments in technology-driven, classroom-based innovative assessment practices. It uses a framework organized around cognitive constructs, assessment functionality, and automaticity to review the technological developments of innovative assessments and identify how they have been advanced to meet researcher and practitioner needs. Specifically, technology such as artificial intelligence (AI), learning analytics, or virtual reality (VR), among others, plays six significant roles: (1) simulating real-world problems; (2) eliciting, capturing and representing complex performance; (3) analyzing processing data; (4) visualizing performance and providing actionable interpretation; (5) transforming scores into actionable information; and (6) ease human effort.

This section also discusses the potential of technology-based innovative assessments in supporting the effectiveness, equity, and feasibility of complex STEM teaching and learning. Attention is also paid to the need for continued design and development research to explore this potential. The authors also provide recommendations for refocused, extensive training for how humans can use AI to solve workplace challenges. They stress the need for supporting and improving human (e.g., teacher) judgment skills around decision-making in classrooms, especially under conditions of time stress and uncertainty, and apply deliberation, ethics, and practical knowledge. They also suggest that teachers need to learn how to leverage the quantity and quality of information provided by AI-driven assessment technologies.

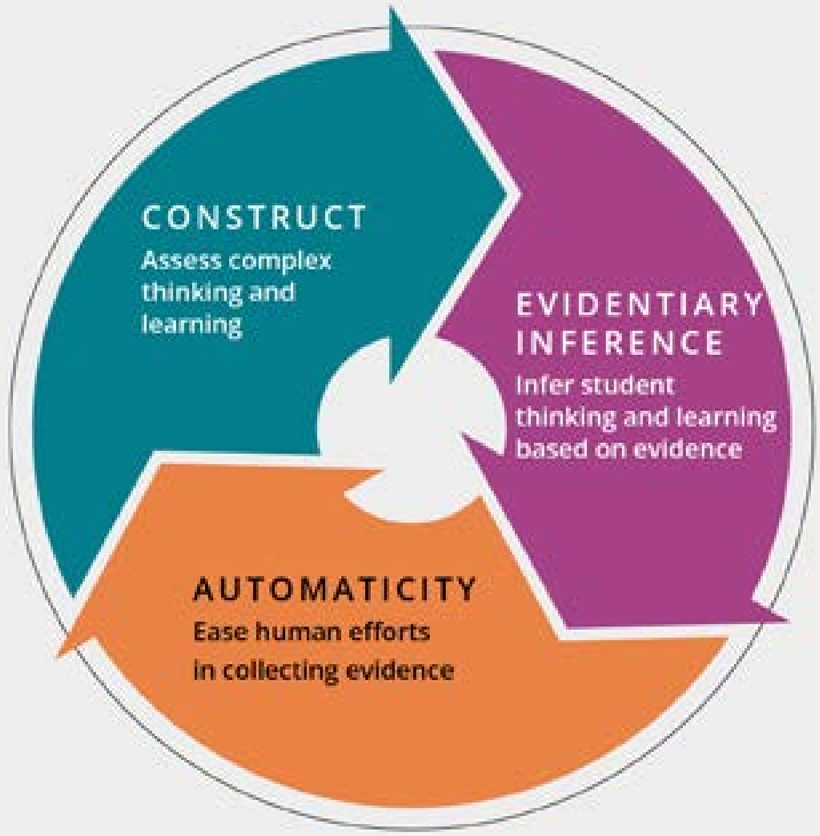
Introduction

To meet the challenges of educating children and preparing a future workforce, instructors should be able to track students' thinking and progression toward learning goals in a timely fashion, using classroom-based assessments that are reliable, valid, and feasible. Not surprisingly, technology has shaped how these formative assessments are designed, administered, and analyzed (Lawless & Pellegrino, 2007; Pellegrino, 2010; Zhai & Pellegrino, in press). A large body of research has adopted technologies such as games, VR, AI, and learning analytics for developing classroom-based assessments (Bennett, 2018; Jiao & Lissitz, 2020; Linn et al., 2014; Zhai, Yin et al., 2020). Technology-based assessments that target complex constructs, augment the assessment functionality of evidentiary reasoning, and assist teachers in classroom assessment practices can be considered *innovative assessment practices* (Zhai, 2021b). While previous research and policy work shows some of the potential of technology-based innovative assessments, synthesizing these developments and identifying the role of technology, as technology bridges our observations, interpretations, and understanding of students' cognition (NRC, 2001), is critical to forward the field both theoretically and practically.

In this section, we present an overview of critical developments in technology-driven, classroom-based innovative assessment practices, using a framework organized around *construct*, *assessment functionality* (i.e., evidentiary inferences), and *automaticity* (Zhai, 2021b; Zhai, Haudek et al., 2020). Innovative assessment practices are expected to leverage cutting-edge technologies to elicit performance, collect evidence, analyze data, and make evidentiary inferences about students' critical and complex learning outcomes. Such activity will target complex constructs—the attributes of examinees that a test intends to assess. Meanwhile, innovative assessment practices advance an assessment's functionality and feature an evidentiary inference process. These accessible and feasible assessment practices are expected to free teachers from time-consuming tasks (e.g., grading students' responses) and save costs. This section intends to review the technological developments of innovative assessments from the lens of these three dimensions to identify how assessments have been advanced using technologies to meet societal needs (**Figure 4**).

Specifically, we aim to (a) identify the critical roles that emerging technologies play in assessment practices and (b) show how these technologies have advanced assessment practices by targeting complex constructs, increasing assessment functionality, and easing human effort. Of particular focus will be emerging computational tools, such as educational learning analytics and AI, and their role in the form and functionality of classroom-based assessment. To demonstrate the effort that has been made to facilitate the roles of technology in assessment practices, we will highlight one project from the NSF DRK-12 portfolio. This section concludes with how this review of prior work and emerging technologies come together to point toward future directions in classroom-based assessment development, implementation, and research.

Figure 4. Framework for technology-based innovative assessment (graphic used with permission of Xiaoming Zhai, author)



Assessing Complex Constructs

Classroom instructional reforms in STEM are driven in part by the need for a skilled workforce who is capable of meeting the rapid changes in the nature of work and using knowledge to enhance the nation's competitiveness, security, and leadership in the world (NASEM, 2019; NRC, 2012b). The skill sets needed for this workforce to succeed are multifaceted and are increasingly challenging to assess in the classroom—the more complex the skill sets, the more challenging it is to develop quality assessments that capture the constructs in a classroom setting (Furtak, 2017). In science education specifically, and STEM education more generally, it is widely acknowledged that such skill sets can be developed most effectively through meaningful learning integrating three dimensions—science and engineering practices (SEPs), crosscutting concepts, and disciplinary core ideas (DCIs; NRC, 2012b). In such three-dimensional learning environments, students are offered opportunities to engage in scientific modeling, argumentation, and explanations by using scientific knowledge to solve problems and figure out solutions. This three-dimensional learning represents the new vision of STEM learning that has significant potential to achieve societal goals and the needs of the workforce market.

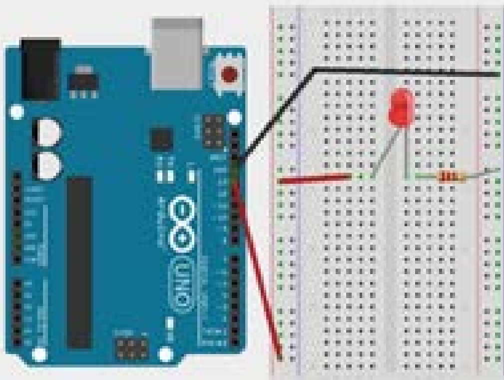
However, implementing the three-dimensional vision is not without challenges, particularly with regard to classroom assessment practices. Three-dimensional learning is a complex construct and thus is a demanding assessment practice (Pellegrino, 2013), particularly in classroom settings where teachers have limited resources to elicit, interpret, and make instructional decisions. Integrated with science and engineering practices, three-dimensional learning requires tasks that engage students in performing the practices and provide evidence to infer their thinking (Harris et al., 2019). Furthermore, the *Framework for K–12 Science Education* (NRC, 2012b) requires the assessment tasks to be able to track students' learning progression—the developmental attributes of children. These challenging activities require assessments that allow teachers to capture multimodal channels of a student's learning progression and associated thinking during problem-solving, as well as to what degree their skill sets can be successfully transferred to solve real-world problems in new scenarios (Duschl et al., 2007; Smith et al., 2018; Zhai et al., 2021). For these innovative assessment practices, emerging technologies have played a critical role.

Role 1. Simulating real-world problems to elicit students' thinking and knowledge-in-use

Students' ability to use scientific knowledge to solve real-world problems can be best assessed if they are presented with tasks that authentically mirror real-world problems. It is well known that authentic tasks have a basis in the real world and thus must be shown contextualized in authentic environments (Pellegrino, 2012). While the prior (tasks) is relatively feasible to accomplish, the latter (environments) can often be challenging without technologies. The development of technologies provides avenues to serve either as materials for learning or as assessment tools to engage students in solving real-world problems (Gao et al., 2022a; Gao et al., 2022b). A salient example of the first category is makerspace, which leverages technologies to engage students in hands-on activities to solve problems and has shown promise as an assessment tool (Tang et al., 2020).

Yin et al. (2022) used Arduino, an open-source physical computing platform with software to assess students' computational thinking. They first identified a set of real-world problems and employed Arduino to design the assessment tasks. Arduino can read signals such as light, pressure, or messages via sensors and convert the signals into outputs (e.g., activating a motor or lighting an LED). Focusing on the electricity unit, they engaged students in maker activities using both hardware and software design and implementation to solve several real-world problems. **Figure 5** shows an activity in which students were involved in designing SOS signals on LED using Arduino. Yin et al. (2022) suggest that the Arduino-based performance assessments yielded valid assessment conclusions about students' computational thinking and provided advanced process information about students' ability to use knowledge to solve problems. In this example, technology serves as learning materials to bridge science and engineering learning with real-world problems. Thus, this technology provided a vehicle for engaging in real-world tasks that elicited knowledge-in-use aligned with complex constructs using a cost-effective, scalable platform.

Figure 5. Maker activity using Arduino to design SOS signals (Yin et al., 2022; used with permission)



```
// Activity 2: SOS signals

int led = 13;                  // led connected to pin 13

void setup() {
  pinMode(led, OUTPUT);        // led is an output
}

void loop() {
// creating the first "S": 3 short flashes of light
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);

// creating "O": 3 long flashes of light
    digitalWrite(led, HIGH); delay(1000);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(1000);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(1000);
    digitalWrite(led, LOW);  delay(500);

// creating the second "S": 3 short flashes of light
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);
    digitalWrite(led, HIGH); delay(500);
    digitalWrite(led, LOW);  delay(500);

    delay(2000);                         // wait for 2 seconds, then
  repeat
}
```

Assessments that target complex constructs are usually constrained by the scenarios that students can experience. Validity of assessments can be significantly reduced if students receive a limited or low-fidelity experience that authentic problems require, denying students an opportunity to apply their knowledge. Recent developments in learning technologies such as simulations, VR, and augmented reality (AR) can create rich, expansive, authentic scenarios, allowing students to experience the tasks in depth. In their study, Chang et al. (2020) used VR to create a geological park for peer assessments of earth science (**Figure 6**). Compared to their peers in the control group, the students using peer assessment with VR increased their environmental awareness and achieved increased earth science learning gains. This study also suggests that technologies that created authentic scenarios used in classroom assessments could engage students' perceptive channels in ways that increased learning motivation.

Figure 6. VR environment in which students can create a project for peer assessment (Chang et al., 2020; used with permission of Morris Jong)

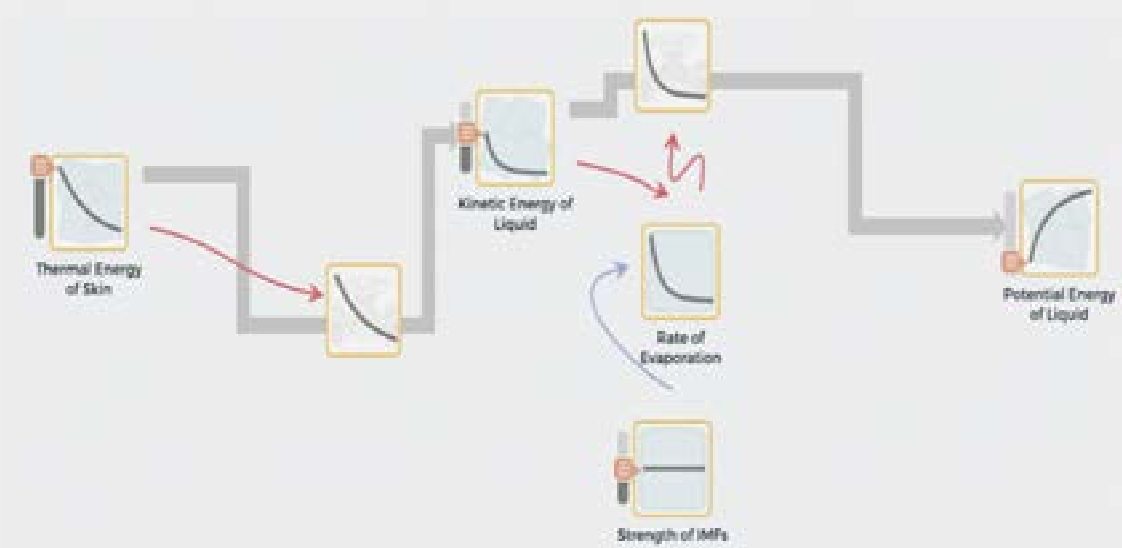
Role 2. Eliciting, capturing, and representing complex performance

Due to the complexity of STEM learning that involves science and engineering practices, it is vital to use technological tools to support students in representing ideas so that assessors can feasibly capture their otherwise invisible thinking. Traditional representation channels limit students' ability to express their ideas and thus create challenges in the assessment process (Zhai, 2022). This is particularly important since each communication modality has the potential to reveal unique information about a student's understanding. For example, research has shown that students' drawn models and written descriptions of models are only weakly associated even though they are used to explain the same phenomena (Zhai, He, & Krajcik, 2022). While normal paper-pencil tests using handwriting limit equitable opportunities for students with lower writing proficiency, technologies such as simulations and drawing tools provide additional venues for students to represent their ideas (Donnelly-Hermosillo et al., 2020; Smith et al., 2018).

In their study, Chang et al. (2014) developed Chemation, a drawing tool that allows students, regardless of their writing or hand drawing proficiency, to construct 2D molecular models.

With Chemation, students could change molecular components using an atom palette. Their findings suggest that modeling tools revealed insights into students' thinking processes while crafting explanatory models for phenomena. To further support classroom assessment practice where students may test, revise, and debug models, Damelin et al. (2017) developed SageModeler, a semi-quantitative system modeling application. Figure 7 shows a model that a student developed using SageModeler to explain evaporative cooling (Bowers et al., 2022). Students could test their models by varying the input and observing how the outputs and mediators change dependently.In their research, technology serves as a representation tool to help students express their ideas and assist teachers in both eliciting and capturing complex performance.

Figure 7. Student-drawn model for evaporative cooling using SageModeler (Bowers et al., 2022; used with permission of the National Science Teachers Association)



Model-based explanation of evaporative cooling.

An example of a student's final model along with a transcription of their explanation of evaporative cooling based on their model. Student explanation: This model shows how evaporation causes us to feel colder when we are wet. The thermal energy of our skin makes the liquid warmer (increasing its temperature and kinetic energy). As the liquid particles have more kinetic energy, they will move faster and begin to break free from the liquid and evaporate. Once the liquids are free from these intermolecular bonds, they become gas molecules with high potential energy. Liquids with a higher IMF evaporate slower because it requires more energy to overcome the intermolecular bonds. I included some "feedback" relationships because the more energy there is in the skin, the faster the liquid will heat up and the more kinetic energy in the liquid, the faster the liquid will evapo-rate.

Making Evidentiary Inference

Assessment practices are challenging partially because the constructs to be assessed are (mentally) internal to examinees and invisible to assessors, and thus valid assessment conclusions must be made upon assessors' inferences based on indirect evidence. Using traditional formative assessment practices (e.g., multiple choice tests), the common evidence available to teachers is raw scores, which provide limited, at best, direct actionable information for instruction (Shepard et al., 2018). Teachers must interpret raw scores and use their interpretations to make instructional decisions (Bennett, 2011; Black & Wiliam, 1998). This is a challenging task, especially in classroom settings where teachers need to make immediate instructional decisions with limited time and resources. Historically, there has been a tension between using quick, scalable assessment techniques that provided little actionable data and richer, more robust elicitation approaches that were too time-consuming and required logistically challenging physical manipulatives. Recent developments in the above-mentioned interactive computer environments, learning analytics, and machine learning provide opportunities to upend this historic tension, offering real-time or near-real-time customized feedback to students and information to teachers to advance and enhance inclusive STEM learning (Viberg et al., 2018; Zhai, 2021a).

Role 3. Analyzing process data to shed light on cognitive processes

Learning analytics is an interdisciplinary area made prominent by the availability of large-scale log file data generated in open online courses (Ferguson, 2012; Siemens, 2013; Viberg et al., 2018). As instruction was virtualized and almost 100% contained within a computer system, new possibilities to more holistically capture and assess user behavior opened up. Gašević et al. (2022) classified learning analytics as FOR assessments (i.e., learning analytics as forms of assessment) or OF assessment (i.e., applying learning analytics to answer questions about assessment practices), indicating the strong links between learning analytics and assessment practices. Using learning analytics, educators can trace users' interaction with technologies during problem-solving to shed light on students' learning process and uncover theories accountable for students' problem-solving behaviors (Gao et al., 2022a; Gao et al., 2022b). Learning analytics has shown the potential to promote evidentiary inferences by identifying attributes and student behavioral patterns, which can be used to support instruction and adaptive learning.

A study conducted by Xing et al. (2015) analyzed students' log file data while engaging in virtual math teams with Georgebra (VMTwG) to assess their collaborative learning. VMTwG is an online platform that allows students to collaboratively solve math problems via a whiteboard with a set of geometry tools and to communicate with each group member via chat messages. Xing and his colleagues collected and analyzed five types of log file data: VMT awareness logs (e.g., erasing chat messages); Georgebra logs (e.g., actions of constructing geometry artifacts such as adding a point, updating a segment); system logs (i.e., order of user actions); chat message content; and whiteboard logs (e.g., drag or draw using the tools provided). They first applied cluster analysis to identify collaborative groups and then developed tools to visualize group behaviors. **Figure 8** shows the performance of the three groups. Within each group, students' performance levels are visualized, then by clicking each group, a radar graph will pop up in a teacher dashboard to show every individual's performance on six dimensions. These learning analytics outcomes uncovered students' performance patterns that helped inform instruction.

An ongoing challenge with these forms of more distal data collection, it is often difficult to ascertain both what is transpiring away from the computer—either through behaviors (e.g., gestures) that don't involve computer input or dimensions of behavior not readily captured through keyboard or mouse (e.g., facial expressions). However, the same audio and video technologies that facilitate online Web conferencing can be leveraged to expand the modalities through which students can express their knowledge and also provide parallel affective data that can help flesh out a deeper understanding of student states (e.g., Grafsgaard et al., 2014; Zakaria et al., 2021). Leveraging multimodal trace data in analysis often improves models utilizing AI-driven analytic techniques (Wiggins et al., 2018).

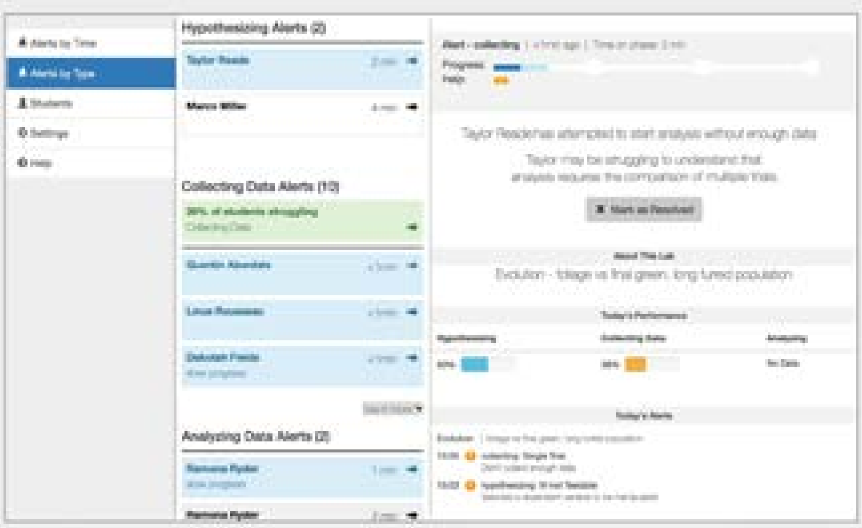
Figure 8. Collaborative learning outcomes from learning analytics (Xing et al., 2015; licensed under CC BY 3.0)

Role 4. Dashboard design to visualize performance and facilitate actionable interpretation

Even learning analytics-driven classroom assessments are often constrained by the lack of timely and usable information provided to teachers in an organized manner for informed instructional decision-making. A recent development in learning analytics has spurred better teacher dashboard design, an outcome that potentially could improve the informed classroom instructional decision-making process (Matcha et al., 2020; Schwendimann et al., 2017). It is critical that a useful dashboard design is human-centered, in that the information displayed is timely, valid, and useful (Ahn et al., 2019; Latif et al., in press).

As an example, Inq-ITS is a virtual lab that supports students in conducting scientific inquiry. Using this lab, students can ask questions or form hypotheses about a phenomenon; carry out simulated investigations and collect data; analyze and interpret data; and write explanations using claims, evidence, and reasoning (Gobert et al., 2015; Gobert et al., 2013). To facilitate effective learning, Dickler et al. (2021) developed Inq-Blotter, a teacher-alerting dashboard that facilitates teachers' timely support for students' scientific inquiry practices (**Figure 9**). Teachers could organize the alerts according to the types of practice or recency (based on how long students have been struggling). The dashboard also allows teachers to view a list of students' performance outcomes. Flexible design is provided so that teachers can configure what to display (e.g., diagnostic information) and how to display the alerts triggered by automated analysis of this information. The dashboard also provides a voice recording function to capture teachers' guidelines (see **Figure 9**). Dickler et al. (2021) found that teacher supports provided by Inq-Blotter facilitated students' scientific inquiry performance.

Figure 9. Screenshot of an Inq-Blotter alert with the microphone voice recording feature (Gobert et al., 2023; used with permission). Clicking on a student's name presents the teacher with diagnostic information about the student's difficulty.

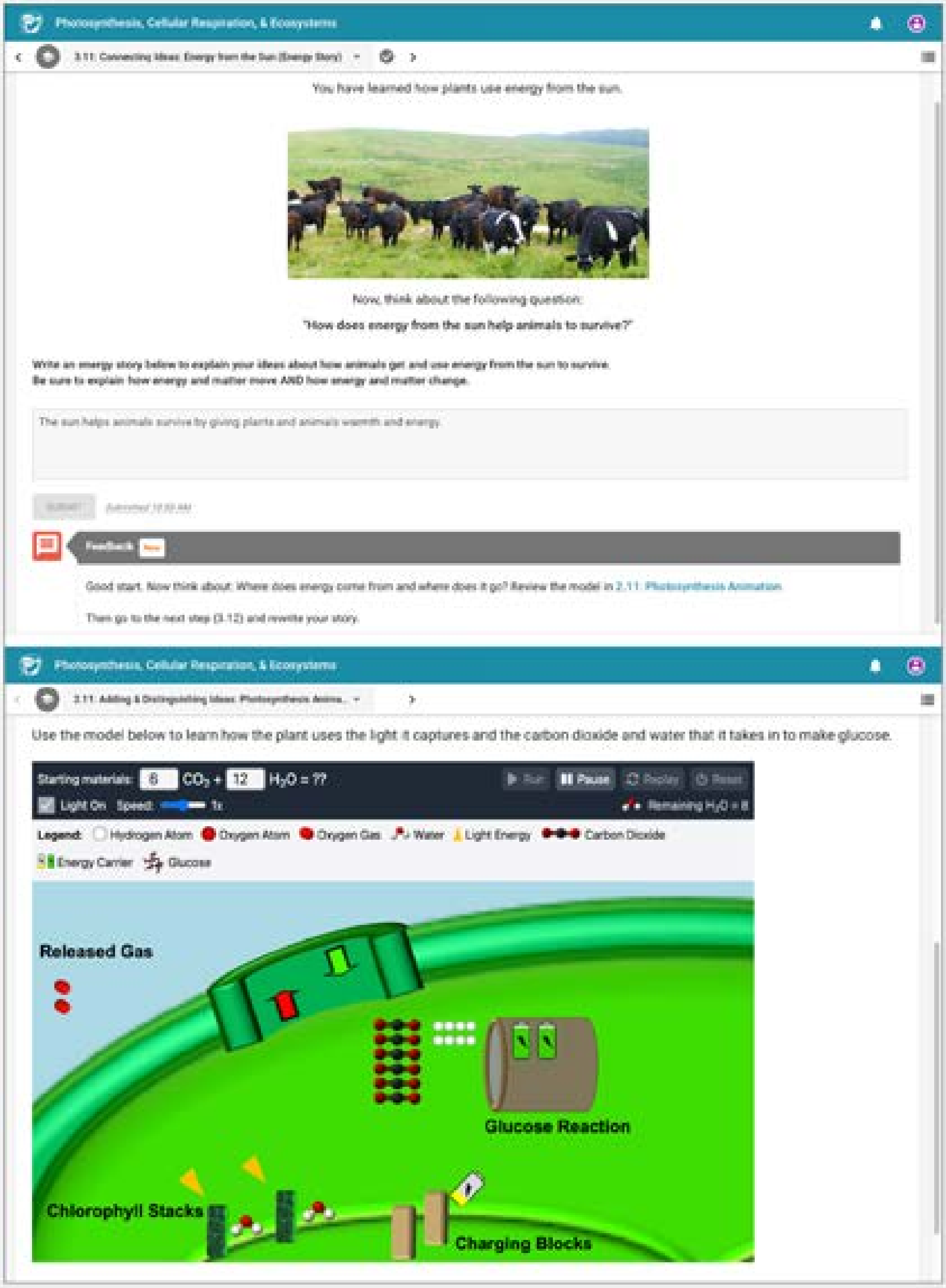


Role 5. Transforming student scores into comprehensible and actionable information

The availability of automatically generated student scores on more complex behavioral data spurs research on the design of effective learning and teaching guidance. Research has shown significant progress in automatic scoring (Zhai, Haudek et al., 2020), which provides the opportunity to individualize learning supports so that students with diverse backgrounds can learn STEM at their own pace. Using automatic scoring technologies, teachers orchestrating learning activities in the classroom are able to have more anchors to understand individual students' needs and provide customized support (Linn et al., in press; Zhai, 2021a). Together with learning analytics, these technological developments have prompted personalized learning supports, transforming scores into actionable knowledge both directly for students and directly to teachers in support of students.

The Web-based Inquiry Science Environment (WISE) team has developed automated learning and teaching guidance, leveraging automatic scoring according to the knowledge integration framework (Gerard & Linn, 2016). **Figure 10** shows a student's written responses to the Energy Story task in a photosynthesis project. The computer returned a rubric score of 2 of 5, indicating the student's response lacked any normative links among energy ideas. Based on this score, the computer provided direct learning guidance, asking the student to revisit a dynamic visualization of energy transfer and transformation inside a chloroplast, shown in the image below, and then to revise their response. The guidance includes four major components: students' current ideas, a question prompting students to think of the missing or non-normative idea, a suggestion to revisit relevant evidence, and a prompt for the revision of students' responses based on evidence. At the same time, students might also be reminded to consult their teacher for further learning guidance if their revisions received scores lower than 3 (out of 5). The same information can be structured and delivered to teachers to inform their actions supporting classroom learning. In this type of classroom, technologies have played a role in transforming assessment scores into actionable information for both learning and teaching.

Figure 10. Courtesy of Web-based Inquiry Science Environment [(https://wise.berkeley.edu)](https://wise.berkeley.edu)



Automaticity

A fundamental purpose of any technology, including learning technologies, is to pursue automaticity to ease human burden and reduce the level of effort and cost. Current approaches to assessment are typically time-consuming and often perceived by teachers as tedious work that involves substantial efforts in design, development, administration, scoring, and reporting. However, it is also an area where technologies have shown substantial potential (Linn et al., 2014).

Role 6. Easing human efforts and reducing cost

Earlier waves of improving assessment through technology focused on scoring (e.g., OCR for multiple-choice bubble sheets). The advancement of speed and accuracy of historically rich but time-consuming assessments opens the possibility of using these assessments at scale to provide novel insights. By moving assessments into digital (i.e., computer-based) form, the speed of embedded tools moves assessments that were formerly considered distal summative assessments into the realm of (near) real-time proximal formative tools. The move of assessments online has also meant that interactivity can be built into item response, opening the door to a wider range of problem types and providing more engaging problem spaces for students (Grover et al., 2022). This interactivity can thus create more of a dialogue with students, providing real-time feedback as they engage with the items, thus coming closer to the goals of formative assessment. Even for more traditional forms of assessment, due to recent developments in machine learning, automatic scoring of multiple-choice and constructed responses has begun moving from simply reporting scores to the deeper analytic interpretation of patterns of inter- and intra-student responses (Beggrow et al., 2014). In this dimension, recent research has focused on improving the technologies' fundamental features of automaticity beyond simply improved speed and accuracy to leverage assessment at scale to provide unique insight into student learning.

Effective STEM teaching and learning relies on timely feedback through classroom assessments. Traditional multiple-choice items can be feasibly graded using computerized technologies, but the new performance-based assessments for next generation STEM learning that involves science and engineering practices (e.g., modeling, argumentation, and explanations) are challenging to score automatically. The recent development in machine learning opens up the window from which researchers can create algorithmic models using expert-graded student responses to automatically score students' constructed responses and artifacts (Liu et al., 2016; Zhai, Haudek et al., 2020; Zhai, Yin et al., 2020). For example, researchers have applied natural language processing to automatically score students' constructed responses of conceptual understanding (Nehm & Haertig, 2012); explanations (Maestrales et al., 2021); and argumentations (Lee et al., 2021; Wilson et al., in press; Zhai, Haudek, & Ma, 2023).

With the goal of broadening the modalities with which students can express their ideas, researchers have explored technologies to automatically score students' visual representations. Smith et al. (2018) employed a topology-based algorithm to assess students' drawn responses in a tablet-based digital science notebook.This approach first defines a set of possible relations between objects representing scientific phenomena as *near*, *far*, and *contains*. Rules were also applied to limit the relations between certain components so that the resulting outputs could be bounded. Their items were fairly structured, and students were required to use the components provided to map a scientific model by placing and orienting them relative to each other (see **Figure 11**). The interactive icon-based modeling, while constraining expression, lowered student's cognitive load, didn't depend on the drawing skill of the student, and enhanced the automated scoring.

As a fast-growing field, novel methods are constantly being developed to automatically score student responses, including students' hand-drawn models. Zhai, He, & Krajcik (2022) applied the RestNet-50 convolutional neural networks to grade students' drawn models and yielded robust accuracy (**Figure 12**). They collected middle school students' performance on modeling tasks that targeted the NGSS expectations via a computer-based tool, through which students could freely draw models either using provided components (e.g., arrows) or not. While the prior work employed computers to collect students' responses, a recent study by Wang et al. (2022) focused on paper-pencil drawing (**Figure 13**). They asked students to draw models about optical phenomena (e.g., refraction) and employed 2D convolutional neural networks to score high school student responses, receiving satisfactory results.

Figure 11. The modeling tool used by Smith et al. (2018)[5]

5 © 2019 IEEE. Reprinted, with permission, from IEEE Transactions on Learning Technologies.

Figure 12. "Red dye diffusion" item screenshot (left), response interface (right), and a student response (bottom; Zhai, Haudek, & Ma, 2023; reprinted with permission of Springer Nature)

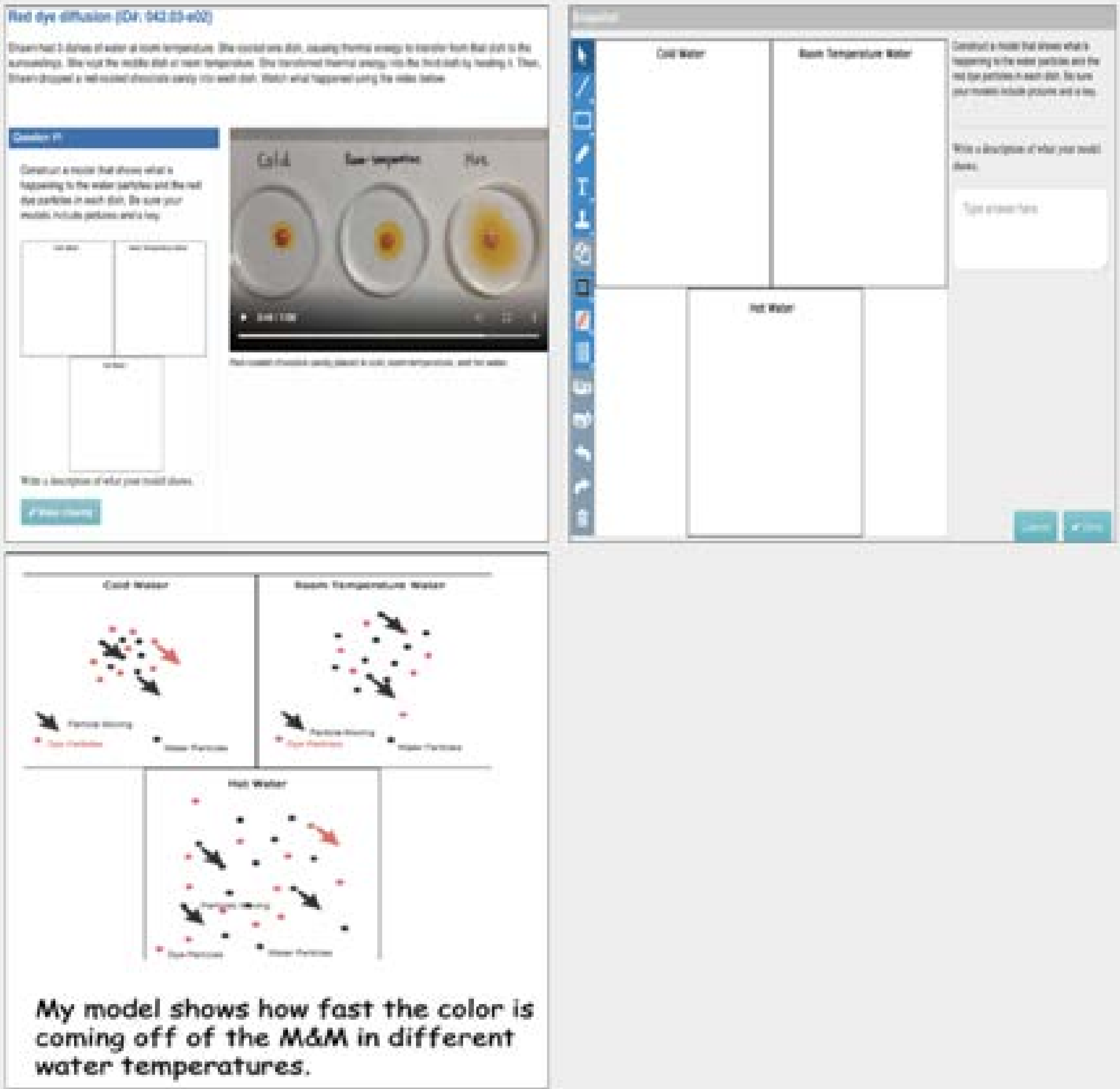
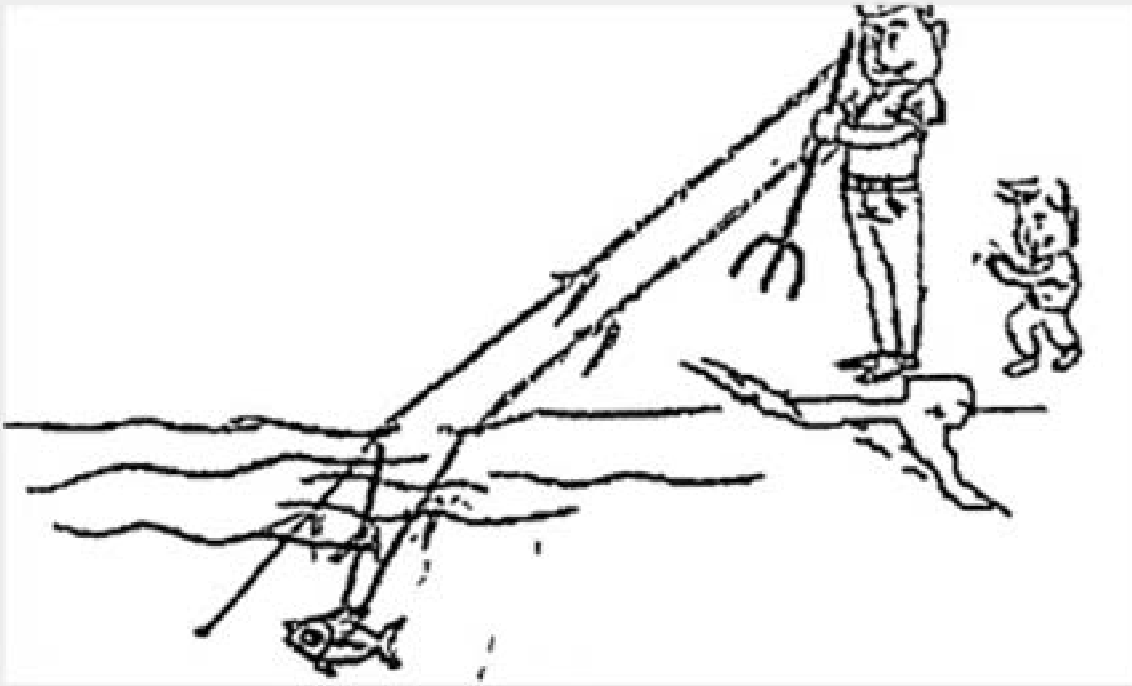


Figure 13. Student paper-pencil drawn model on refraction phenomena (used with permission of Xiaoming Zhai, author)



The automatic scoring was also applied in computer science (CS) learning. An NSF-funded exploratory effort draws on Grover's (2021) formative assessment framework for K–12 CS as well as research on code comprehension, Parson's problems, subgoal labeling, and multiple-choice question (MCQ) assessments to go beyond MCQ items and design innovative technology-driven assessments that are auto-gradable, interactive, and engaging (Grover, et al., 2022). The project leverages the affordances of technology on an assessment platform, Edfinity (also developed with NSF support), to create innovative item types. Drag-drop items such as Parsons problems (**Figure 14**) or match-the-column or one where a student drags 'subgoal labels' from a palette to the appropriate section of code could be both more engaging and better for formative feedback. They designed these and item types, such as hotspot (**Figures 15** and **16**) or point-and-click, where instead of providing images of grids as options for where a robot will end up on a grid after code execution, a student simply clicks on that square (**Figure 17**). Such a format is more intuitive and lowers cognitive load. The project has also involved creating an item bank that targets granular learning goals and known misconceptions documented in CS research to help teachers with formative feedback on targeted topics and concepts.

Figure 14. Existing MCQ problems changed to Parson's Problems (Grover et al., 2022; used with permission)

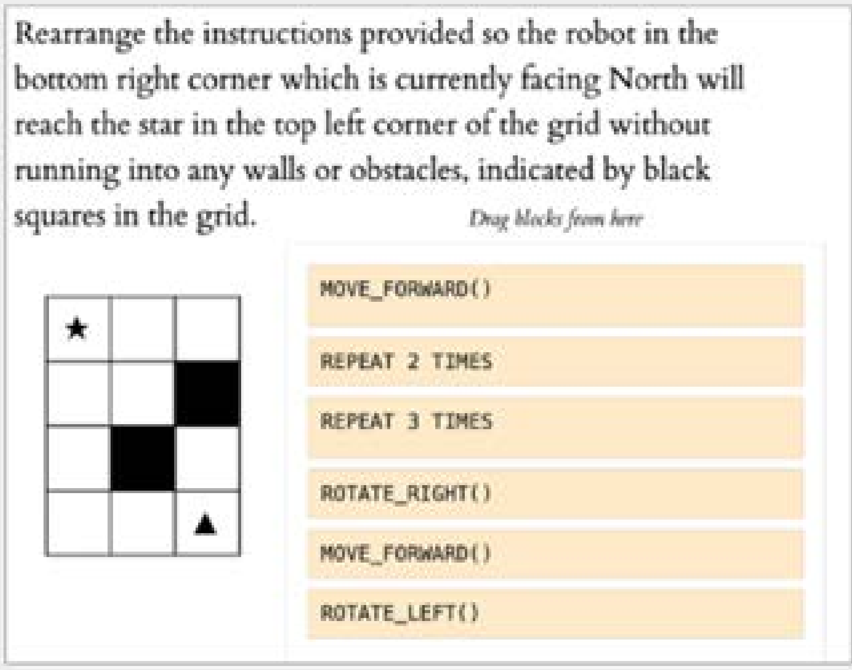


Figure 15. Block-based code snippets converted to a hotspot interaction problem requiring subgoal labeling (Grover et al., 2022; used with permission)

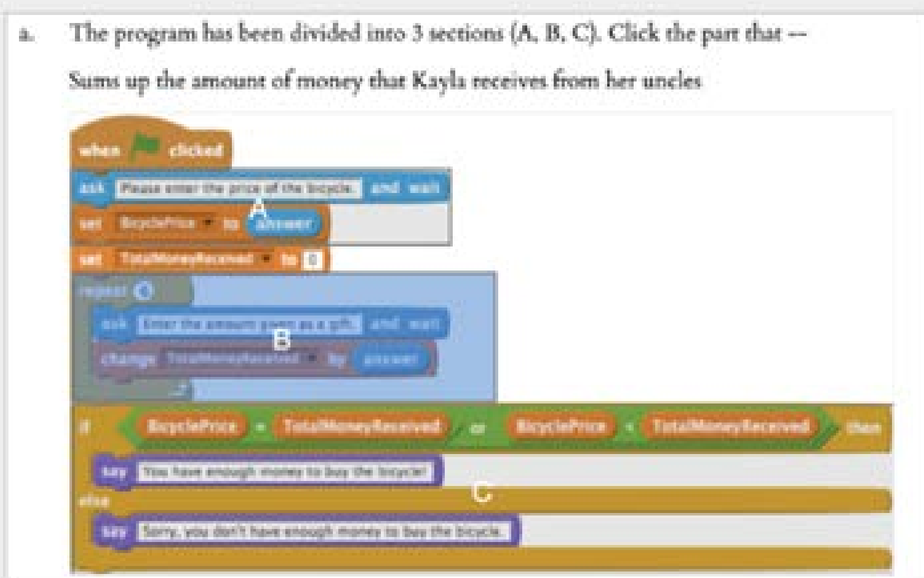


Figure 16. An MCQ item from CTt assessment (Wiebe et al., 2019) adapted into a point-and-click item, which is more intuitive and has a lower cognitive load (image courtesy of Professor Marcos Román-González)

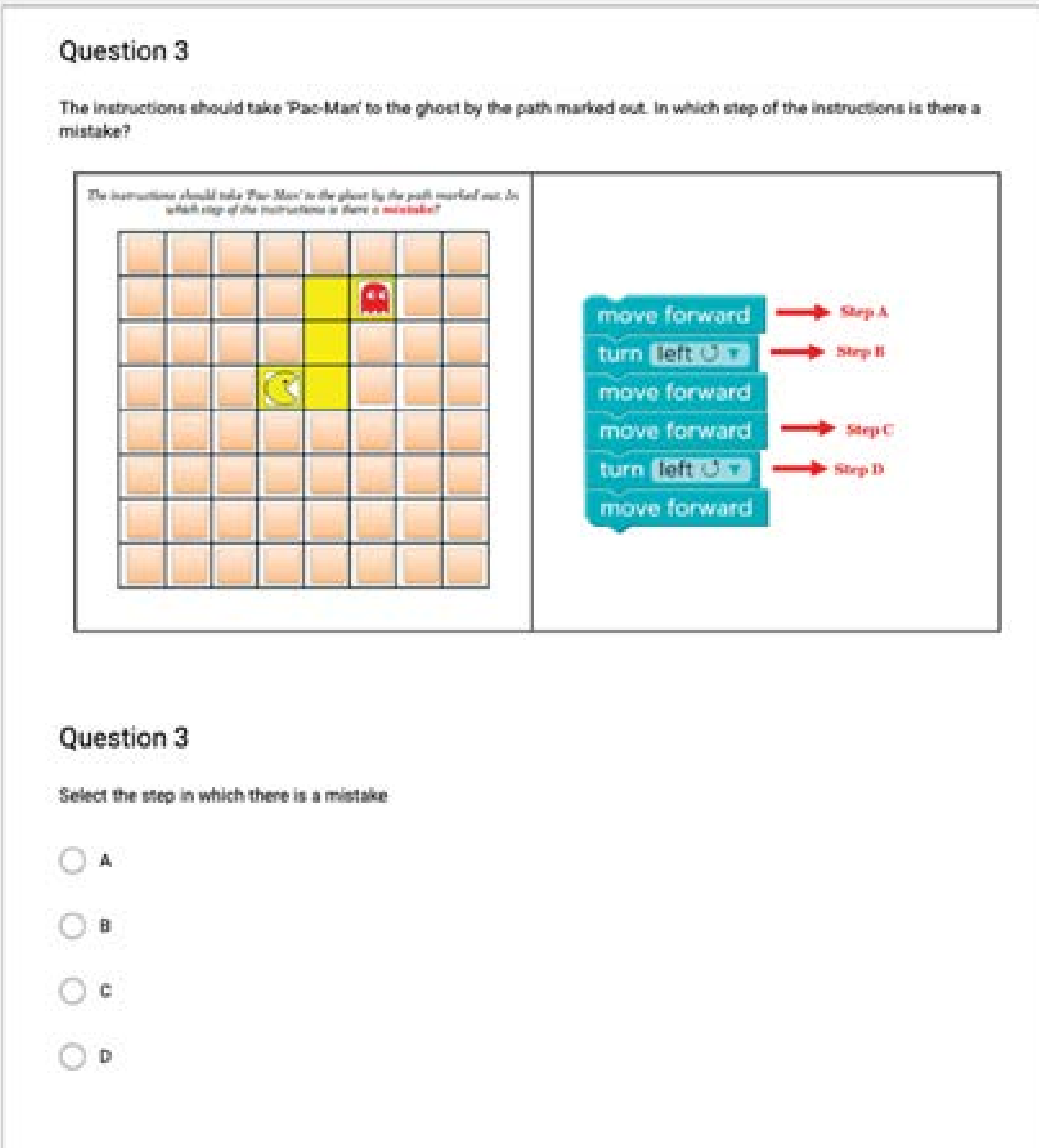
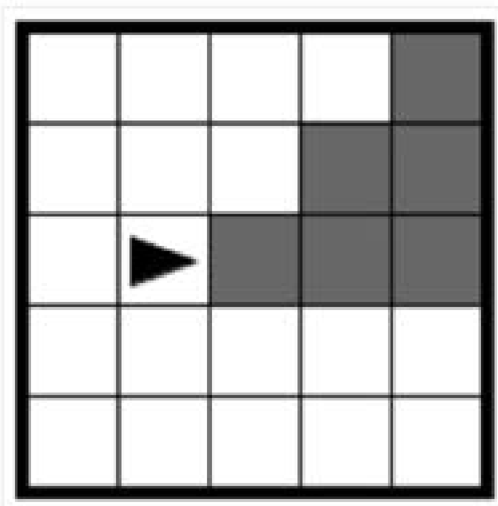


Figure 17. An AP CS Principles sample MCQ question changed to a point-and-click item (Grover et al., 2022; used with permission)

b. Click the square the robot will occupy once the code segment is finished running.

```
IF (CAN_MOVE ( left ))
{
 ROTATE_LEFT ()
 MOVE_FORWARD ()
}
IF (CAN_MOVE ( left ))
{
 ROTATE_LEFT ()
 MOVE_FORWARD ()
}
IF (CAN_MOVE ( left ))
{
 ROTATE_LEFT ()
 MOVE_FORWARD ()
}
```



Portfolio of NSF DRK-12 Project

Supporting Instructional Decision Making: The Potential of An Automatically Scored Three-dimensional Assessment System (PASTA) (# 2101104, 2100964, 2101166, 2101112). The *PASTA* project aims to tackle the challenges of engaging students in science assessment practices that integrate three dimensions of scientific knowledge – science and engineering practice, disciplinary core ideas (DCIs), and crosscutting concepts. Led by scholars from the University of Georgia, Michigan State University, the University of Illinois at Chicago, and WestEd, the project team collaboratively works to develop valid and robust AI and machine learning algorithms to automatically assess students' three-dimensional performance. To use the assessment information in a meaningful way, the team further focuses on developing instructional strategies to support teachers' assessment practices. By using the assessments,automatic scoring system and instructional strategies,teachers will develop pedagogical content knowledge (PCK) to implement 3D science assessment tasks.

This project builds on prior projects in which members of the team had developed three-dimensional assessments aligned with the Next Generation Science Standards (led by Christopher Harris, Joseph Krajcik, and James Pellegrino), as well as AI technologies that can automatically score students' constructed responses (led by Xiaoming Zhai). The project has three goals: (a) develop automatically generated student reports (AutoRs) for three-dimensional (3D) science assessments to assist middle school teachers in noticing, attending to, and interpreting information in ongoing classroom teaching, and (b) develop effective instructional strategies to support teachers' PCK and to improve teachers' use of AutoRs to make effective decisions for instructional moves, and (c) examine the overall effectiveness of AutoRs and instructional strategies to support teachers' decision-making and student 3D learning.

Automatically scored 3D assessments would have a significant impact on teachers' instructional decision-making only if teachers have the PCK of how to make use of the assessment information, the transformation of different knowledge (i.e., 3D science knowledge, knowledge of how students learn science, knowledge of pedagogy and classroom experience) that allows teachers to develop the most useful and powerful forms of representation (i.e., powerful phenomena, tasks, analogies, illustrations, examples, explanations, demonstrations) to support learners at a particular age and experience (Magnusson et al., 1999; Shulman, 1987). This is because assessment itself cannot inform teachers how to transform the knowledge gained into practice (Siegel & Wissehr, 2011), and is particularly true with the novel knowledge-in-use 3D assessments that integrate SEPs, crosscutting concepts, and DCIs that students need to use to make sense of complex and compelling phenomena and problems. Using 3D assessments for instruction creates both cognitive and pedagogical challenges for science teachers, as described below (NRC, 2014).

The first challenge for teachers is how to interpret student responses to 3D assessments. The interpretation demands a more intensive cognitive load and richer knowledge base for teachers than that of traditional science assessments (Furtak, 2017), as 3D assessments are usually performance-based constructed responses that involve multiple factors to mirror authentic scenarios. One potential solution to reduce teachers' cognitive load is to develop and provide teachers with organized assessment reports so that teachers can better notice, attend to, and interpret student responses. However, prior studies suggested that teachers tend to gain limited useful information about student thinking and instructional implications from current forms of student reports (von Aufschnaiter & Alonzo, 2018). Because it is unclear how to format reports to help teachers make immediate instructional decisions to effectively promote student knowledge-in-use (NRC, 2012a), this project will explore teachers' cognitive processes of acquisition and use of student performance information. To this end, the team will develop and refine AutoRs to provide robust and instructionally relevant information to teachers so that they, in turn, can interpret student knowledge-in-use performance.

A second challenge for teachers is how to effectively transform student performance on 3D assessments into effective instruction. Effective instructional moves require not only the ability to interpret assessment results but also create actionable knowledge (Bennett, 2018). If teachers have limited PCK to transform their interpretation of student performance into meaningful 3D instructional activities, 3D assessments might still end up with a limited impact. Therefore, essential instructional strategies such as pedagogical scaffolds might help teachers effectively use information from 3D assessments. This project will develop instructional strategies to help teachers make instructional decisions to effectively promote 3D learning. The project is well-aligned with the DRK-12 Teaching Strand as it catalyzes research in STEM teacher development—the need to help science teachers throughout the nation make sense of 3D assessment information to promote learning targeting the Next Generation Science Standards (NGSS Lead States, 2013) performance expectations.

The collaborative project will unfold over three phases. In the first phase, the team designed AutoRs and explored how to present information on knowledge-in-use assessment tasks in the AutoRs to support teachers' immediate instructional decision-making. The team used ten NGSS-aligned assessment tasks designed by the Next Generation Science Assessment (NGSA) project to assess middle school students' performance expectations on chemical reactions, from which they developed diagnostic rubrics for all assessment tasks used in the project. They employed four types of algorithms to develop scoring models for the assessment tasks, resulting in machine-human agreements above benchmarks.

Having developed a human-centered design framework for the design of AutoRs, the team has worked with middle school teachers through cognitive labs to gain their interpretation and perceptions of the three types of AutoRs. The team also developed a mobile application, AI-Scorer, which can automatically score students' written responses and present AutoRs to teachers as they move about the classroom.

In the second phase, the team will design instructional strategies and study what features of instructional strategies could support teachers' use of AutoRs to make effective instructional decisions. To support teachers' use of AutoRs, the team is developing instructional strategy supports to be used in the classroom setting. The ultimate goal is to improve teachers' PCK of classroom assessment practices with AutoRs. Phase three will pilot test the AutoRs and the instructional strategies in supporting students' science learning and teachers' instructional decision-making. The team will employ a cluster randomized trial study to explore the usability of AutoRs and instructional strategies to support teachers' instructional decisions and students' 3D learning.

Recommendations for Future Research and Practice

This section attempts to articulate six salient roles of technologies in advancing assessment practices to facilitate STEM learning. We reported both the achievements and knowledge we have learned through research and development of technology-driven innovative assessments. The results clearly point to the need for continued design and development research to explore the potential of technology-based innovative assessments in supporting the effectiveness, equity, and feasibility of complex STEM teaching and learning. Compelling research has demonstrated the potential, but research and development work is needed to realize the potential of these tools. Based on this review, we provide the following recommendations.

Both the research presented here and the broader work on the near-term impact of AI on the workforce points more to the need for refocused, extensive training for how humans and AI will work together to solve workplace challenges (Dede et al., 2021). While AI has become increasingly proficient at the computation of assessment results and focused prediction of learning outcomes, there is now a heightened need for supporting and improving human (i.e., teacher) actionable decision-making in classrooms, especially under conditions of time stress and uncertainty. This decision-making by teachers needs to apply deliberation, ethics, and practical knowledge. While master teachers have always employed these skills, they now need to learn how to leverage the quantity and quality of the information provided by AI-driven assessment technologies that they previously had not had access to. Understandably, as technologies move into classrooms, they will take over the most rote tasks, opening the door for teachers to engage in more higher-order formative assessment. As emerging AI-driven tools help reveal more complex constructs of student understanding in STEM (e.g., the *PASTA* project), teachers need support in learning how to interpret these new constructs, how they align with curricula, and how to support students in their learning and skill development in the classroom.

Research needs to continue the nascent foundational work on how new formative, classroom-based assessment technologies are operationalized in the classroom (Ertmer et al., 2012). As both the type and quantity of assessment data evolve, teachers and professional development providers need to work on how best to deploy the reception of data and implementation of teacher actions into differing time frames of the classroom orchestration loop (Roschelle et al., 2013; Sharples, 2013). For example, what data should be delivered in real time during a class period for immediate action versus at the end of the week or term for reflection on future implementations of a unit? Future research will also help inform a better understanding on how to strategize about the distribution and operationalization of real-time assessment data between what is delivered directly to students within computer-based learning environments, what goes to teachers to use as part of their classroom decision-making and orchestration, or some combination thereof.

While AI-driven technologies hold the promise of delivering finer-grained, deeper understanding of individual student learning, it does not, in and of itself, provide solutions to the long-term structural inequalities that have led to unequal opportunities for minoritized populations of students and the systematic misinterpretation of student ability and potential. Both the development and training of analytic models sensitive to a broad range of cultural experiences and expressions of knowledge as evidence, and the training of teachers as to how to

operationalize insights in more emphatic ways will be important starting points for this work. **Section 3** of this report provides a fuller discussion of these issues of equity and justice. Finally, research by teams consisting of both CS and educational researchers will continue to develop new machine learning algorithms, learning analytics, and software development tools. These algorithms will make use of a broader range of data streams, not only coming from student keyboard and mouse interaction with learning environments, but also video, VR, AI at the student and classroom level interpreting biometric and classroom-level movement data, natural language (both voice and text) processing, along with other emerging possibilities.
This work will also continue to refine unsupervised machine learning algorithms that will not need time-consuming human-coded data sets to train on. All of these approaches, inevitably, will bring on new, challenging ethical issues that need to be addressed in parallel with the development of these emerging technologies.

References

Ahn, J., Campos, F., Hays, M., & Digiacomo, D. (2019). Designing in context: Reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics, 6*(2).

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science Education and Technology, 23*(1), 160–182.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational measurement: issues and practice, 37*(4), 7–15.

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

Bowers, J., Damelin, D., Eidin, E., & McIntyre, C. (2022). Keeping cool with sagemodeler. *The Science Teacher, 89*(4), 18–25.

Chang, H.-Y., Quintana, C., & Krajcik, J. (2014). Using drawing technology to assess students' visualizations of chemical reaction processes. *Journal of Science Education and Technology, 23*(3), 355–369.

Chang, S.-C., Hsu, T.-C., & Jong, M. S.-Y. (2020). Integration of the peer assessment approach with a virtual reality design system for learning earth science. *Computers & Education, 146*.

Damelin, D., Krajcik, J. S., Mcintyre, C., & Bielik, T. (2017). Students making systems models. *Science Scope, 40*(5), 78.

Dede, C., Etemadi, A., & Forshaw, T. (2021). *Intelligence augmentation: Upskilling humans to complement AI.* Cambridge.

Dickler, R., Gobert, J., & Sao Pedro, M. (2021). Using innovative methods to explore the potential of an alerting dashboard for science inquiry. *Journal of Learning Analytics, 8*(2), 105–122.

Donnelly-Hermosillo, D. F., Gerard, L. F., & Linn, M. C. (2020). Impact of graph technologies in K–12 science and mathematics education. *Computers & Education, 146*, 103748.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K–8* (Vol. 500). National Academies Press.

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers & Education, 59*(2), 423–435. **https://doi.org/10.1016/j.compedu.2012.02.001**

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education, 101*(5), 854–867.

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022a). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129*, 107142.

Gao, Y., Zhai, X., Bulut, O., Cui, Y., & Sun, X. (2022b). Examining humans' problem-solving styles in technology-rich environments using log file data. *Journal of Intelligence, 10*(3), 38.

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. *Computers in Human Behavior, 134*.

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education, 27*(1), 111–129.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43–57.

Gobert, J. D., Sao Pedro, M.A., Betts, C.G. (2023). An AI-based teacher dashboard to support students' inquiry: Design principles, features, and technological specifications. In N. Lederman, D. Zeidler, & J. Lederman (Eds.), *Handbook of Research on Science Education* (Vol. 3, pp. 1011–1044). Routledge. **https://doi.org/10.4324/9780367855758**

Grafsgaard, J. F., Wiggins, J., Vail, A., K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction.* Association for Computing Machinery. **http://dx.doi.org/10.1145/2663204.2663264**

Grover. S. (2021). Toward a framework for formative assessment of conceptual learning in K–12 computer science classrooms. *Proceedings of the 52nd ACM Technical Symposium on Computing Science Education (SIGCSE'21)*. Association for Computing Machinery.

Grover, S., Carmichael, B., & Venkatasubramaniam, S. (2022, March). Beyond MCQ: Designing engaging, autogradable assessments for supporting teaching & learning in K–12 computer science. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education* (Vol. 2; pp. 1123–1123). **https://dl.acm.org/doi/10.1145/3478432.3499080**

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67.

Jiao, H., & Lissitz, R. (2020). *Application of artificial intelligence to assessment*. Information Age Publishing.

Latif, E., Zhai, X., Amerman, H., & He, X. (in press). Automatic scoring platform for AI-based formative assessment. In Z. X. & K. J. (Eds.), *Uses of Artificial Intelligence in STEM Education*. Oxford University Press.

Lawless, K. A., & Pellegrino, J. W. (2007). Professional development in integrating technology into teaching and learning: Knowns, unknowns, and ways to pursue better questions and answers. *Review of Educational Research, 77*(4), 575–614.

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology, 30*(2), 168–192.

Linn, M. C., Donnelly-Hermosillo, D., & Gerard, L. F. (in press). Synergies between learning technologies and learning sciences: Promoting equitable secondary science education. In N. Lederman, D. Zeidler, & J. Lederman (Eds.), *Handbook of research on science education* (Vol. III). Routledge Press.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science, 344*(6180), 155–156.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Krajcik, J., & Schneider, B. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology, 30*(2), 239–254.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95–132). Springer.

Matcha, W., Uzir, N. A., Gasevic, D., & Pardo, A. (2020). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies, 13*(2), 226–245.

National Academies of Sciences, Engineering, and Medicine,. (2019). *Science and engineering for grades 6–12: Investigation and design at the center*. National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. **https://doi.org/10.17226/10019**

National Research Council. (2012a). *Education for life and work: Developing transferable knowledge and skills in the 21st century* (J. W. Pellegrino & M. L. Hilton, Eds.). National Academies Press.

National Research Council. (2012b). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. The National Academies Press. **https://doi.org/10.17226/18409**

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of Science Education and Technology, 21*(1), 56–73.

NGSS Lead States. (2013). *Next generation science standards: For states, By states*. National Academies Press.

Pellegrino, J. (2010). Technology and formative assessment. *International Encyclopedia of Education,* 42–47. **https://doi.org/10.1016/B978-0-08-044894-7.00700-4**

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831–841.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*(6130), 320–323.

Roschelle, J., Dimitriadis, Y., & Hoppe, U. (2013). Classroom orchestration: Synthesis. *Computers & Education, 69*, 523–526.

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1), 30–41.

Sharples, M. (2013). Shared orchestration within and beyond the classroom. *Computers & Education, 69*, 504–506.

Shepard, L., Penuel, W., & Pellegrino, J. (2018). Classroom assessment principles to support learning and avoid the harms of testing. *Educational Measurement: Issues And Practice, 37*(1), 52–57.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371–391.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380–1400.

Smith, A., Leeman-Munk, S., Shelton, A., Mott, B., Wiebe, E., & Lester, J. (2018). A multimodal assessment framework for integrating student writing and drawing in elementary science learning. *IEEE Transactions on Learning Technologies, 12*(1), 3–15.

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, *148*(C). **https://doi.org/10.1016/j.compedu.2019.103798**.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98–110.

von Aufschnaiter, C., & Alonzo, A. C. (2018). Foundations of formative assessment: Introducing a learning progression to guide preservice physics teachers' video-based interpretation of student thinking. *Applied Measurement in Education, 31*(2), 113–127.

Wang, C., Zhai, X., Shen, J. & Herdliska (2022). Applying machine learning to assess paper-pencil drawings of optics. Presented at the Annual Meeting of the National Association for Research in Science Teaching, Vancouver, BC. NARST.

Wiebe, E., London, J., Aksit, O., Mott, B. W., Boyer, K. E., & Lester, J. C. (2019). Development of a lean computational thinking abilities assessment for middle grades students. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19)* (pp. 456-461). Association for Computing Machinery. **https://doi.org/10.1145/3287324.3287390**

Wiggins, J. B., Kulkarni, M., Min, W., Mott, B., Boyer, K. E., Wiebe, E., J. Lester, J. (2018). User affect and no-match dialogue scenarios: An analysis of facial expression. *Proceedings MA3HMI'18 International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction.* ACM. **https://dl.acm.org/doi/abs/10.1145/3279972.3279979**

Wilson, C., Haudek, K., Osborne, J., Stuhlsatz, M., Cheuk, T., Donovan, B., Bracey, Z., Mercado, M., & Zhai, X. (in press). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*.

Xing, W., Wadholm, R., Petakovic, E., & Goggins, S. (2015). Group learning assessment: Developing a theory-informed analytics. *Journal of Educational Technology & Society, 18*(2), 110–128.

Yin, Y., Khaleghi, S., Hadad, R., & Zhai, X. (2022). Developing effective and accessible activities to improve and assess computational thinking and engineering learning. *Educational Technology Research and Development*, *70*, 951–988. **https://doi.org/10.1007/s11423-022-10097-w**

Zakaria, Z., Tsan, J., Vandenberg, J., Lynch, C. F., Boyer, K. E., & Wiebe, E. N. (2021). Identifying productive conflict during upper elementary students' collaborative programming. *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021* (pp. 261–262). International Society of the Learning Sciences. doi:10.22318/cscl2021.261

Zhai, X. (2021a). Advancing automatic guidance in virtual science inquiry: From ease of use to personalization. *Educational Technology Research and Development, 69*(1), 255–258.

Zhai, X. (2021b). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology, 30*(2), 1–11.

Zhai, X. (2022). Assessing high-school students' modeling performance on Newtonian mechanics. *Journal of Research in Science Teaching*, 1–41. **https://doi.org/10.1002/tea.21758**

Zhai, X., Haudek, K., & Ma, W. (2023). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education, 53*, 405–424. **https://doi.org/10.1007/s11165-022-10062-w**

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459.

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, 1–30.

Zhai, X., Krajcik, J., & Pellegrino, J. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology, 30*(2), 298–312.

Zhai, X., & Pellegrino, J. (in press). Large-scale assessment in science education. In N. G. Lederman, D. L. Zeidler, & J. S. Lederman (Eds.), *Handbook of research on science education* (Vol. III, pp. 1045–1098). Routledge.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111–151.

Suggested Citation

Zhai, X., & Wiebe, E. (2023). Technology-based innovative assessment. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives* (pp. 83-94). Community for Advancing Discovery Research in Education, Education Development Center, Inc.

Summary and Recommendations for Classroom-Based STEM Assessment

The STEM education community is at a critical juncture in how we approach assessment and its role in the teaching and learning process. As emphasized in this report, major developments over the past two decades in STEM education and classroom assessment provide a compelling argument for shifting the focus of assessment toward the classroom where its most powerful benefits can be realized. The increased emphasis on diversity, equity, and inclusion in education puts increased scrutiny on assessment practices and points to the need for more proximal approaches that are contextualized in the places where students learn and sensitive to individual differences. Similarly, emergent technologies, especially those centered on AI technologies, are providing potent tools for teachers to use in classrooms to assist with assessment that can guide more individualized instruction. As a classroom-based activity necessarily impacted by both calls for increased attention to equity and justice as well as powerful, emerging technologies, the need for substantive professional development is once again brought to the forefront.

This report has attempted to highlight what has been learned about STEM classroom assessment based on education research that has emerged over the last 20 years. This work has been guided by two emergent theoretical ideas. First, sociocultural theory has provided an overarching view of how teaching and learning unfolds and how assessment may play a role in guiding this process. In parallel, work on learning progressions and trajectories has provided fine-grain hypothesized pathways of learning for specific disciplinary content areas, and it continues to maintain relevance in emergent STEM subjects. These two central theoretical ideas work with the triadic connection between learning goals and its associated curricular materials, instructional strategies to achieve these learning goals, and classroom-based assessment that helps both in the reflecting back on the success of the chosen methods and looking forward to what should come next and for whom.

This report looks at current research and critical issues regarding classroom-based assessment in the STEM disciplines from a number of perspectives, with the various sections of the report each taking a different lens to classroom-based assessment, while also collectively providing a holistic view of this critical enterprise in STEM classroom teaching and learning. The conclusions reached and recommendations going forward are intended to be useful and usable for multiple audiences, including members of the CADRE community; program directors and project officers at the National Science Foundation; and STEM education researchers, practitioners, and policymakers. It is hoped that the contents of this report will stimulate dialogue among members of these communities that will enhance implementation of effective assessment practices in K–12 STEM education classrooms and the content and execution of in-service and preservice teacher professional learning programs to support effective assessment, while also helping to chart a course for high-priority areas for the next decade for STEM classroom assessment research, development, and implementation to be funded by NSF, other federal agencies, and private foundations.

Charting a Course: Recommendations for STEM Classroom Assessment

Each section of the report represents different and important integrations and perspectives on the knowledge and practice of STEM classroom assessment. On their own, each of the five sections offer recommendations for research, practice, and policy. Collectively, they point to a number of themes of particular note. The recommendations are based on what we currently know as well as what we need to know to make classroom assessment in STEM maximally beneficial for the instructional practices of teachers and the learning outcomes of students. What follows are major recommendations from the five sections of the report.

Recommendations for Connecting Classroom Assessment with Learning Goals and Instruction through Theories of Learning (Section 1)

Models of how learning progresses in the STEM disciplines, often referred to as either "learning progressions" or "learning trajectories," offer a compelling and principled way for developing assessments that align with learning goals and instruction and cohere with learning theory. This alignment is critically important for ensuring that assessments reflect contemporary views on learning in STEM disciplines and that they serve to provide insight into how students' disciplinary knowledge and practices are developing over time with appropriate instruction.

- **Recommendation 1-1.** *As learning progressions continue to be mapped out and empirically validated in the STEM disciplines, it will be important to focus research efforts on their use as a framework for developing and using assessments that inform instructional decision-making.* Learning progressions have tremendous potential to guide the design of classroom-based assessments for formative purposes, and the role they can play in classroom-based assessment remains an important area for researchers.

- **Recommendation 1-2.** *Standards and expectations for STEM proficiency have changed substantially, and assessments for today's STEM classrooms should reflect these contemporary perspectives on learning in the disciplines*. Notably, research and development are needed on specific ways in which classroom-based assessments can be designed to reflect multiple ways of knowing in STEM disciplines and that honor and reflect students' cultural practices and funds of knowledge.

- **Recommendation 1-3.** *More research work is needed to help us better understand the ways teachers can generate meaning from assessment results that will transform students' opportunities to learn.* It is widely recognized that alignment among curriculum, instruction, and assessment is integral to coherent and robust STEM teaching and learning, yet there is still much to be learned about how assessments and assessment results can be used in instructionally supportive ways.

Recommendations for Assessment for Learning (Section 2)

As opposed to assessment *of learning*, which is generally summative in nature, formative assessment entails assessment *for learning.* It is characterized by the effort to collect information about the state of students' thinking and performance so instruction can be tailored to further progress. Learning progressions describe the development of sophistication in students' thinking and performance related to STEM discipline-specific topics. Aligning classroom assessments with learning progressions may improve the relevance and usefulness of classroom-based assessment for informing instruction and guiding future learning opportunities.

- **Recommendation 2-1.** *High-quality assessment development and validation work are needed for classroom-based STEM assessments that can be used to improve teaching and advance learning.* Validation requires rigorous design coupled with multiple sources of evidence supporting assumptions about what the designed tasks are assessing and their capacity to reveal important student differences that are conceptually and instructionally meaningful.

- **Recommendation 2-2.** *Ongoing research is needed about the nature and efficacy of professional learning for supporting teachers to implement formative assessment effectively.* The shift toward assessment for learning often presents conceptual issues for teachers in understanding and implementing the formative assessment process integrated with curriculum and instruction.

- **Recommendation 2-3.** *Development, iterative refinement, and validation of learning progressions across all STEM disciplines should continue and is especially encouraged in emerging areas of technology, engineering, and computer science education.* This work is critical for establishing a continuum within and across grade levels with various implications for classroom STEM assessment.

Recommendations for Equity and Justice in Classroom Assessment of STEM Learning (Section 3)

Equity and justice should be centered in STEM learning environments and in the assessment of STEM learning. For equity and justice to be centered in classroom assessment, it must also be centered in curriculum and instruction so that all three major components of the learning environment can work effectively together. Increasingly, developers and practitioners are taking up equity and justice perspectives in developing or adapting assessments that are responsive to the learners who will be participating in them. To accomplish this, an expansive view of both equity and justice in assessment needs to be undertaken. Researchers and practitioners need to embrace their unique roles as they work on developing assessments and associated professional development materials that are responsive to the learning needs of all students and integral to curriculum and instruction.

- **Recommendation 3-1.** *Contemporary perspectives of equity and justice should be central in STEM classroom assessment design and practice.* It is important to recognize that centering equity and justice in STEM classroom assessment is a transformative act that will require accompanying transformations in curriculum and instruction. Also noteworthy is that more of the projects in the DRK-12 portfolio are taking approaches associated with equity than with justice in assessment. Accordingly, the portfolio should move toward an increased emphasis on incorporating justice in classroom assessment of STEM learning.

- **Recommendation 3-2.** *The DRK-12 program should take a more expansive view of assessment as integral to curriculum and instruction and consider that all three must work together toward new futures in STEM learning.* This view is needed to better understand students' experiences with assessment, their opportunities to engage in meaningful learning, the ways in which they are supported through instruction, and how their ideas and experiences are represented in and taken up in learning.

- **Recommendation 3-3.** *In the development of assessments that center equity and justice, designers should build assessments in concert with educators, learners, and communities.* STEM education practitioners and leaders are ideally positioned to center equity and justice in developing or adapting curriculum materials and assessments and in implementing instructional approaches that build on their relationships with students, families, and communities. In particular, teachers need spaces to reflect upon the role that their classroom assessment plays in assuring that all students have opportunities to engage in STEM learning.

- **Recommendation 3-4.** *Assessment work should embrace a wide and inclusive view of what constitutes historically minoritized populations.* The newly energized work on bringing equity and justice to STEM classrooms needs to embrace goals that encompass a wide range of unique populations, which includes not only racial and ethnic cultural diversity, but also physical and neurodiversity as well. Better understanding of what resources different populations bring to the classroom will help in developing and deploying classroom-based assessments that allow all students to demonstrate their abilities and articulate where and how support is needed.

Recommendations for Teacher Knowledge and Practices for Assessment (Section 4)

It is widely recognized that formative assessment is fundamental to high-quality teaching and learning, yet it is among the most challenging practices to embed in everyday instruction. The promise of classroom-based STEM assessment rests on the extent to which teachers understand the formative use of assessment and can incorporate it into their instructional practice. Accordingly, teacher professional learning for assessment with accompanying resources and supports to transform professional learning into practice are paramount.

- **Recommendation 4-1.** *New assessment development projects should attend to the knowledge and practices teachers need support in, and design assessment systems with features to support professional learning and uptake in practice.* Closely related is identifying high-leverage strategies that can bring about change in classroom-based assessment practices at a large scale.
- **Recommendation 4-2.** *Disciplinary-specific models of student cognition should be explicitly articulated in ways that can be examined apart from the assessment tools designed to elicit thinking, and so that teachers can understand them and make use of them.* Cognitive models can serve to support teachers, curriculum and assessment developers, and researchers to share a vision of student learning that is grounded in research in a particular STEM domain.
- **Recommendation 4-3.** *Assessment to be used in classroom practice should elicit student ideas and reasoning for the purpose of monitoring and advancing individual and collective learning.* To accomplish this, prompts should be designed to allow for diverse ways in which students can demonstrate learning and performance. This is needed so that teachers can gain insights to support further learning rather than only judging students' ideas in relation to conventional ideas and models that professional communities use.
- **Recommendation 4-4.** *STEM practitioners should undertake cycles of designing, implementing, and reflecting with colleagues to improve assessment practice over time and develop shared strategies for dealing with the practical challenges of designing and implementing assessment.* Collaborative planning and conversations can help move teachers away from using traditional or narrow assessment techniques and formats toward trying new ways for using assessment to support STEM learning.

Recommendations for Technology-Based Innovative Assessment (Section 5)

The development of technology-based innovative assessments is already underway and expanding the boundaries for what assessment can look like in STEM classrooms. Recent progress in a range of areas, including AI, real-time assessment, digital technologies, and VR, combined with new techniques from measurement science and data analytics are playing critical roles in enabling innovative assessment practice. Emerging technologies enable real-world simulations of phenomena and problems; support the eliciting, capturing and representing of complex performance; efficiently analyze both response and response process data; aid in displaying and interpreting performance; transform individual and collective data into actionable information; and stand to reduce many of the practical challenges of formative assessment.

- **Recommendation 5-1.** *Those responsible for the design and development of technology-based innovative STEM assessments should continue to explore how technology can work with teachers and students to support and improve the effectiveness, equity, and feasibility of complex STEM teaching and learning.* Compelling research and development is underway, but more work is needed on how assessment technologies can be operationalized for formative-focused use in classrooms.

- **Recommendation 5-2.** *Cross-disciplinary teams of experts from diverse fields, including computer science, STEM education, psychology, assessment, and ethics (among others), are needed for envisioning, designing, and ensuring the development of AI-driven classroom-based assessment systems that benefit all students and teachers.* While there is tremendous promise with AI-driven technologies, there are also many potential pitfalls and new, challenging issues regarding privacy, fairness, equity, and access that will need to be addressed in parallel with the development of these emerging technologies.

- **Recommendation 5-3.** *Research is needed on the distribution and operationalization of real-time assessment data with regards to what is delivered directly to students, what goes to teachers to use as part of their classroom decision-making and orchestration, or some combination thereof.* As both the type and quantity of assessment data evolve, there is a need to better understand how data can be organized, represented, and deployed with next-step guidance to students and teachers within the differing time frames of the classroom assessment orchestration loop.

Priority Areas with Implications for Research, Practice, and Policy

The curriculum-instruction-assessment triad, discussed in the Introduction of the report, is one way to understand how intertwined assessment is with all aspects of classroom-based teaching and learning, as highlighted in many sections of this report. Similarly, other sections of the report point out how assessment research, development, and deployment need to be attentive to the issues surrounding equity and justice and the potential of emerging technologies. Below we put forth five priority areas for research, practice, and policy.

Anchoring the work of assessment in STEM classrooms. By its very nature, work on classroom-based assessment cannot exist in isolation of other key elements of classroom practices. That is, assessment work needs to be done *in service* of instruction, which, in turn, should be driven by standards-based curriculum. This work has to be sensitive to the diverse set of cultural resources, ways of knowing, and ways of expressing oneself that students bring to the classroom. Moreover, technologies need to be designed and utilized to help solve the challenges of classroom-based assessment. To accomplish this, expertise needs to be drawn not only from different disciplinary areas (e.g., computer science, psychology, education), but also from instruction taking into account contemporary perspectives on classroom activity and the students and teachers in today's classrooms.

Assessment design for integrated STEM knowledge and proficiency. Research on learning progressions and learning trajectories points both to the unique, disciplinary-specific knowledge students need and the pathways they follow to develop these understandings. However, the increasing emphasis on integrated STEM also has led to discussion of progressions and trajectories and their associated assessments that need to cut across traditional disciplinary boundaries. If there is one thing all of the research on developing progressions and trajectories has shown us, it is how much more work is still needed within and across disciplines. For instance, development and validation of assessments in areas such as data literacy, computational thinking, or engineering design need contexts rooted in science, mathematics, and technology. Accordingly, assessment meant for interdisciplinary learning needs to measure integrated proficiency with two or more disciplines. This work needs to be done in a way that helps uncover new, generalizable knowledge about STEM learning, but that does not result in a loss of coherent and actionable information for teachers.

Addressing challenges of curricular coherence and aligned assessment. The distributed nature of information and knowledge in the digital age has resulted in curricular resources being drawn from multiple sources that have undergone different levels of review and alignment to national and state standards. Thus, providing coherence between curriculum and assessment for teachers and students in classrooms across a district or state can pose a substantial challenge for the design, selection, and use of assessment, especially at the classroom level. Coherence can only be achieved if the curricular and assessment resources are aligned to the same learning goals and if that alignment holds not only within classrooms but across classrooms at school, district, and state levels.

Building an expanded and inclusive view for meeting the needs of student populations underserved by current assessment models and practices. While the current work in equity and justice has taken up this challenge, it will be important to continue identifying and understanding important but understudied populations (e.g., the neurodiverse), and how research with these populations can be generalized to help formulate guideposts and practices applicable more broadly. A strong approach would be to utilize the Principles of Universal Design for Learning (UDL; Hitchcock, Meyer, & Rose, 2002) to guide the design of curriculum, instruction, and assessment, while providing for multiple means of engagement, representation, and action and expression. This approach has gained currency in the last two decades, especially in new disciplines such as CS, as they align well with emergent calls for equity and inclusion. Curriculum materials designed to promote equity and justice are developed in partnership with teachers, students, and communities using phenomena and problems located in place. In addition, just as recent important work has examined the intersectionality of gender and race/ethnicity, research on understudied populations, such as the neurodiverse, will need to look at the unique characteristics that emerge at the intersections.

Leveraging emerging technologies to unlock the full potential of classroom-based assessment. While there is a strong tendency to look to technologies to simply automate practices already in place, the power and potential of emerging technologies allow us to consider exciting new ways to design and deploy classroom-based assessment. By starting with the aspirational goal of instructionally informative assessment, we can look to technologies to provide ways of achieving what has heretofore not been scalable, along with instructional insights that were previously opaque to teachers, students, and researchers.

The priorities mentioned above can only be fully realized through reciprocal partnerships involving STEM education practitioners and stakeholders. STEM teachers, for example, will need to become integral partners with researchers and developers to co-design and implement technology-enhanced classroom-based assessment tools that accurately reflect the knowledge and abilities of all students in their classrooms. With the collaborative effort of teachers, researchers, developers, and other relevant stakeholders including school district STEM leaders among others, we can come to better understand what teachers need and when they need it. As with the design and deployment of the assessments themselves, technological improvements may point to novel approaches to this challenge.

In the Introduction, we noted that the NSF DRK-12 program has long held the following stance: "For assessment to be a driving knowledge engine that moves STEM education forward it must be integrated with systems of learning and teaching, with specific attention paid to the needs of practitioner communities and how assessments would be used in formal education settings" (NSF, 2020). The major sections of this report, individually and collectively, affirm the wisdom and significance of this statement. Substantial progress has been made in pursuing this integration in STEM classrooms, but many critical issues remain within and across the STEM disciplines. An ambitious, multidisciplinary agenda of research, development, and implementation is needed to fully reap the benefits that can accrue from welldesigned and appropriately implemented assessment tools and practices for the STEM classroom.

References

National Science Foundation. (2020). Discovery Research PreK-12 (DRK-12) Program Solicitation (NSF 20-572). **https://www.nsf.gov/pubs/2020/nsf20572/nsf20572.htm**

Appendix: A Sampling of DRK-12 Classroom-Based Assessment Projects

This report features several DRK-12 projects to illustrate the types of classroom-based assessment research funded by the National Science Foundation (NSF). Listed below is a more extensive sampling of DRK-12 projects with award periods ending on or after October 1, 2018. They are listed in order of project start date and include an excerpt of the abstract, with the full abstract accessible via the award number link. To view a complete list of projects with the keyword "assessment," visit **https://go.edc.org/DRK12assessment**.

Collaborative Research: Designing Assessments in Physical Science Across Three Dimensions

Principal Investigator: Joseph Krajcik, Michigan State University
Award #**1316908** | Award Period: 9/01/2013-8/31/2019

Principal Investigator: Christopher Harris, SRI International
Award #**1316903** | Award Period: 9/01/2013-11/30/2018

Principal Investigator: James Pellegrino, University of Illinois at Chicago
Award #**1316874** | Award Period: 9/01/2013-5/31/2019

Principal Investigator: Christopher Harris, WestEd
Award #**1903103** | Award Period: 9/01/2018-8/31/2019

Abstract: This is a collaborative proposal among the University of Illinois at Chicago, Michigan State University, and SRI International to develop, test, and analyze sets of technology-supported diagnostic classroom assessments for middle school (grades 6-8) physical science. Assessments are aligned with the performance assessment and evidence-centered design methodologies suggested in the Framework for K–12 Science Education (NRC, 2012). The study focuses on the development of new measures of learning that take into account the interdependence of science content and practice. Two disciplinary core ideas—Matter and its Interactions, and Energy—and two scientific and engineering practices—Constructing Explanations and Designing Solutions, and Developing and Using Models—are used for this purpose.

Improving Formative Assessment Practices: Using Learning Trajectories to Develop Resources that Support Teacher Instructional Practice and Student Learning in CMP2

Principal Investigator: Alison Castro Superfine, University of Illinois at Chicago
Award #**1316736** | Award Period: 10/01/2013-9/30/2019

Abstract: The overarching goal of this project is to develop innovative instructional resources and professional development to support middle grades teachers in meeting the challenges set by college- and career-ready standards for students' learning of algebra. This 4-year project includes three major components: (1) development and empirical testing of learning trajectories for linear functions and linear equations, (2) collaborations with teachers of Connected Mathematics Project 2 (CMP2) to create and test a set of instructional resources focused on formative assessment processes, and (3) iterative refinement of a professional development model for engaging teachers with the instructional resources in ways that optimize students' learning of algebra. The professional development activities provide opportunities for teachers to develop specialized content knowledge of learning trajectories for linear functions and equations in algebra, processes for interpreting students' performances with respect to those trajectories and providing feedback and additional instructional activities based on "where" the student is with respect to the overall learning trajectory.

DIMEs: Immersing Teachers and Students in Virtual Engineering Internships

Principal Investigator: Jacqueline Barber, University of California-Berkeley
Award #**1417939** | Award Period: 9/01/2014-8/31/2019

Abstract: The Next Generation Science Standards (NGSS) outline the science competencies students should demonstrate through their K–12 years and represent a commitment to integrate engineering design into the structure of science education. However, achieving this new ideal of teaching and learning will require new curricular and pedagogical supports for teachers as well as new and time-efficient assessment methods. This project will provide such curricular and pedagogical support by developing and evaluating teacher-ready curricular Digital Internship Modules for Engineering (DIMEs). DIMES will be designed to support middle school

science teachers in providing students with experiences that require students to use engineering design practices and science understanding to solve a real-world problem, thereby promoting a robust understanding of science and engineering, and motivating students to increased interest in science and engineering. The modules will also assess students' ability to apply their science knowledge in solving the engineering problem, thereby providing teachers with actionable data about the depth of their students' science and engineering understanding.

Investigating How to Enhance Scientific Argumentation through Automated Feedback in the Context of Two High School Earth Science Curriculum Units

Principal Investigator: Ou Liu, Educational Testing Service
Award #**1418019** | Award Period: 9/01/2014-8/31/2019

Abstract: With the current emphasis on learning science by actively engaging in the practices of science, and the call for integration of instruction and assessment; new resources, models, and technologies are being developed to improve K–12 science learning. Student assessment has become a nationwide educational priority due, in part, to the need for relevant and timely data that inform teachers, administrators, researchers, and the public about how all students perform and think while learning science. This project responds to the need for technology-enhanced assessments that promote the critical practice of scientific argumentation—making and explaining a claim from evidence about a scientific question and critically evaluating sources of uncertainty in the claim. It will investigate how to enhance this practice through automated scoring and immediate feedback in the context of two high school curriculum units—climate change and fresh-water availability—in schools with diverse student populations. The project will apply advanced automated scoring tools to students' written scientific arguments, provide individual students with customized feedback, and teachers with class-level information to assist them with improving scientific argumentation.

Playing with the Data: Developing Digital Supports for Middle School Science Teachers Using Game-Based Formative Assessment

Principal Investigator: James Diamond, Education Development Center
Award #**1503255** | Award Period: 7/01/2015-6/30/2019

Abstract: This project will use cycles of design-based research to build new knowledge about how to facilitate teachers' interpretation and use of digital game-based formative assessment data. The research will also inform the revision and expansion of Playfully, an existing, online data-reporting dashboard that can be used with multiple digital games. The project is a collaboration between researchers at Education Development Center Inc.'s Center for Children and Technology (EDC|CCT) and the assessment and game development teams at GlassLab. The research and development teams will engage in a three-year partnership with 60 middle-grade science teachers working in diverse school settings in different parts of the country. The aim of the project is to refine an online formative assessment platform that utilizes data

from a video game designed to teach argumentation at the middle school level. It provides rigorous research on the design features of data tools and associated materials available to teachers to inform their ongoing instruction (i.e., formative assessment tools) when using game-based platforms.

SimScientists Games: Development of Simulation-Based Game Designs to Enhance Formative Assessment and Deep Science Learning in Middle School

Principal Investigator: Matt Silberglitt, WestEd
Award #**1503481** | Award Period: 8/01/2015-7/31/2020

Abstract: This project will focus on understanding how educational games, designed according to research-based learning and assessment design principles, can better assess and promote students' science knowledge, application of science process skills, and motivation and engagement in learning. The project will develop a new genre of games to serve as formative assessment resources designed to collect evidence of science learning during gameplay, provide feedback and coaching in the form of hints, and reinforce middle grade (6th-8th) students' life science concepts and investigation practices about ecosystems described in the Next Generation Science Standards (NGSS) (Achieve, 2013). The games will build on the designs of the simulation-based, curriculum-embedded assessments developed in previous NSF-funded efforts, which include student progress reports and reflection activities that allow teachers to provide feedback to students and adjust instruction.

Developing Formative Assessment Tools and Routines for Additive Reasoning

Principal Investigator: Caroline Ebby, University of Pennsylvania
Award #**1620888** | Award Period: 9/01/2016-2/29/2020

This design and development project is an expansion of the Ongoing Assessment Project (OGAP), an established model for research-based formative assessment in grades 3-8, to the early elementary grades. OGAP brings together two powerful ideas in mathematics education - formative assessment and research-based learning trajectories - to enhance teacher knowledge, instructional practices, and student learning. Building on a proven track record of success with this model, the current project will translate findings from research on student learning of early number, addition, and subtraction into tools and routines that teachers can use to formatively assess their students' understanding on a regular basis and develop targeted instructional responses. The project involves a development component focused on producing and field-testing new resources (including frameworks, item banks, pre-assessments and professional development materials) and a research component designed to improve the implementation of these resources in school settings. The materials that are developed from this project will help teachers be able to more precisely assess student understanding in the major mathematical work of grades K-2 in order to better meet the needs of diverse learners.

Supporting Chemistry Teachers to Assess and Foster Chemical Thinking

Principal Investigator: Hannah Sevian, University of Massachusetts Boston
Award #**1621228** | Award Period: 9/01/2016-8/31/2022

Abstract: The fundamental purpose of this project is to develop, implement, and study a professional development (PD) model for improving chemistry teachers' formative assessment practices to foster teaching focused on chemical thinking. The PD model seeks to refocus and enhance teachers' abilities to notice, interpret, and respond to students' ideas. Building on previous exploratory work through which a Chemical Thinking Framework was developed, the proposed effort will work with 8th-12th grade teachers in Boston Public Schools and the New England Region to assist them (a) to recognize tools that are useful in eliciting students' chemical thinking, and adapt or design formative assessments; (b) to make sense of students' chemical thinking based on data collected using formative assessments that elicit students' thinking; and (c) to strategize responsive actions that better foster learning chemistry.

Supporting Teacher Practice to Facilitate and Assess Oral Scientific Argumentation: Embedding a Real-Time Assessment of Speaking and Listening into an Argumentation-Rich Curriculum

Principal Investigator: Bryan Henderson, Arizona State University
Award #**1621496** | Award Period: 9/01/2016-8/31/2022

Principal Investigator: Eric Greenwald, University of California-Berkeley
Award #**1621441** | Award Period: 9/01/2016-8/31/2022

Abstract: The fundamental purpose of this project is to support teacher practice and professional learning around oral scientific argumentation in order to improve the quality of this practice in classrooms. To achieve this purpose, the project will examine the validity of a new technology-based formative assessment tool for classroom argumentation—"Diagnosing the Argumentation Levels of Groups" (DiALoG)—for which psychometric validation work has been conducted in a laboratory setting. The DiALoG assessment tool allows teachers to document classroom talk and display scores across multiple dimensions—both intrapersonal and interpersonal—for formative assessment purposes. The project will work with 6th-8th grade science teachers to monitor and support argumentation through real-time formative assessment data generated by the DiALoG instrument.

Development of the Electronic Test of Early Numeracy

Principal Investigator: Arthur Baroody, University of Illinois at Urbana-Champaign
Award #**1621470** | Award Period: 9/15/2016-8/31/2023

Abstract: The project will develop and refine an electronic Test of Early Numeracy (e-TEN) in English and Spanish, focused on number and operations. The assessment will incorporate a learning trajectory that describes students' development of the understanding of number. The electronic assessment will allow for the test to adapt to students' responses and incorporate games to increase children's engagement with the tasks. These features take advantage of the electronic format. The achievement test will be designed to be efficient, user-friendly, affordable, and accessible for a variety of learning environments and a broad age range (3 to 8 years old). The overarching goal of the assessment design is to create a measure that is more accurate, more accessible to a wider range of children, and easier to administer than existing measures. This project is funded by the Discovery Research Pre-K-12 Program, which funds research and development of STEM innovations and approaches in assessment, teaching and learning.

Collaborative Research: Modeling Assessment to Enhance Teaching and Learning

Principal Investigator: Mark Wilson, University of California-Berkeley
Award #**1621265** | Award Period: 12/01/2016-2/28/2022

Principal Investigator: Richard Lehrer, Vanderbilt University
Award #**1621088** | Award Period: 12/01/2016-8/31/2023

Abstract: This project will modify an existing assessment system (BEAR Assessment System) to provide ongoing, instructionally productive evidence to teachers about student learning and to link student work products and formative assessments with summative assessments in models that generate useful estimates of student growth. To design and test the assessment system, researchers will study teacher integration of assessment tools with instruction via classroom observations, video records, and interviews. Feedback from teachers and observations of their assessment practices will inform revisions to the assessment system. Multiple iterations will focus on how best to represent and display assessment results for tracking individual and group learning. Researchers will investigate new psychometric models that link information from student classroom work, responses to formative assessments, and summative evaluations to provide more reliable estimates of student learning.

Collaborative Research: Developing & Evaluating Assessments of Problem Solving

Principal Investigator: Jonathan Bostic, Bowling Green State University
Award #**1720646** | Award Period: 9/01/2017-8/31/2023

Principal Investigator: Toni May-Sondergeld, Drexel University
Award #**1720661** | Award Period: 9/01/2017-8/31/2022

Abstract: Unfortunately, most problem-solving assessments are generally framed by a set of mathematics expectations that differ from state standards. Thus, results from those assessments are disconnected from the mathematics content that students learn in the classroom. Previously, this research team has built problem-solving measures for grades 6-8, which address this gap in framing and generates meaningful, valid, and reliable scores, and do not have unintended negative consequences on students. The current project, titled Developing and Evaluating Assessments of Problem Solving (DEAP), builds upon the team's prior work by creating problem-solving measures for grades 3-5. The elementary assessments will be connected to the middle-grades assessments and will be available for use by school districts, researchers, and other education professionals seeking to effectively measure children's problem solving.

Examining an Innovative Approach to Supporting Science Teachers Practice towards Three-Dimensional Learning Goals through Adapting Classroom Assessment Tasks

Principal Investigator: William Penuel, University of Colorado at Boulder
Award #**1748757** | Award Period: 9/01/2017-12/31/2020

Abstract: The project will conduct a two-iteration comparative research study that explores differences in outcomes between science teachers who use only the support tools as resources to adapt classroom tasks versus those who use the tools as part of a two-day professional development experience. The outcomes include changes in teachers' instructional vision related to three-dimensional science learning as measured by a teacher vision survey and teacher interviews, and improvements in the quality of teachers' adapted assessment tasks as determined by a rubric-based analysis and observations of the professional development sessions. Data analysis will follow a conjecture-mapping strategy to compare changes over time within and across groups in order to evaluate hypothesized links between design features of the tools and professional development models, mediating processes, and outcomes. Findings of this study will provide evidence to inform instructional leaders and designs for science teacher preparation, and lead to future development and research in this area.

Collaborative Research: Improving Multi-dimensional Assessment and Instruction: Building and Sustaining Elementary Science Teachers? Capacity Through Learning Communities

Principal Investigator: Carla Strickland, University of Chicago
Award #**1813938** | Award Period: 7/01/2018-6/30/2023

Principal Investigator: James Pellegrino, University of Illinois at Chicago
Award #**1813737** | Award Period: 7/01/2018-6/30/2023

Abstract: This is an Early-Stage Design and Development collaborative effort to better understand how to build and sustain the capacity of elementary science teachers in grades 3-5 to instruct and formatively assess students in ways that are aligned with contemporary science education frameworks and standards. To achieve this goal, the project will use classroom-based science assessment as a focus around which to build teacher capacity in science instruction and three-dimensional learning in science. The three dimensions will include disciplinary core ideas, science and engineering practices, and crosscutting concepts. These dimensions are described in the Framework for K–12 Science Education (National Research Council; NRC, 2012), and the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). The project will work closely with teachers to co-develop usable assessments and rubrics and help them to learn about three-dimensional assessment and instruction. Also, the project will work with teachers to test the developed assessments in diverse settings, and to create an active, online community of practice.

Usable Measures of Teacher Understanding: Exploring Diagnostic Models and Topic Analysis as Tools for Assessing Proportional Reasoning for Teaching

Principal Investigator: Yasemin Copur-Gencturk, University of Southern California
Award #**1813760** | Award Period: 9/01/2018-8/31/2023

Abstract: The overall goal of this project is to pursue a potentially transformative approach to the assessment of teacher proportional knowledge by developing a measure that is well aligned with the content and skills taught in various PD programs. This instrument will be based on a new approach that builds on emerging psychometric models. Specifically, diagnostic classification models (DCMs) will be utilized to diagnose teachers' learning during a PD program as well as employed to identify the progression in teachers' learning. Statistical topic models (STMs) will be used to look for patterns of understanding that emerge from open-ended responses and provide natural-language insight into teachers' reasoning. A final version of the assessment will be constructed for a national sample based on the results from the predictive validity stage, and this version will be tested with teachers who participate in various types of PD programs targeting proportional reasoning.

Collaborative Research: Articulating a Transformative Approach for Designing Tasks that Measure Young Learners' Developing Proficiencies in Integrated Science and Literacy

Principal Investigator: Alison K. Billman, University of California-Berkeley
Award #**1853951** | Award Period: 12/15/2018-11/30/2019

Principal Investigator: Christopher Harris, WestEd
Award #**1853927** | Award Period: 12/15/2018-11/30/2019

Principal Investigator: Daisy Rutstein, SRI International
Award #**1853923** | Award Period: 12/15/2018-11/30/2019

Abstract: SRI International, University of California-Berkeley (Lawrence Hall of Science), and WestEd will join efforts to articulate a potentially transformative approach for designing new kinds of classroom-based, three-dimensional assessment tasks that measure first graders' proficiencies in integrated science and literacy learning. The main goal of this study will be to conduct exploratory-design work to produce both the design approach and the early-stage tasks that are critical inputs for creating a program of research and development to more fully develop a suite of innovative assessment tasks for the early grades. Specific goals of the effort will be:

(1) to iteratively develop and refine a design approach that enables assessment designers to develop Next Generation Science Standards (NGSS)-aligned tasks and rubrics that include a literacy component for the early grades; (2) to use this design approach to create two exemplar assessment tasks that are feasible for classroom use; and (3) to collect initial evidence that informs the promise of the design approach.

Human Variance and Assessment for Learning Implications for Diverse Learners of STEM, A National Conference

Principal Investigator: Edmund Gordon, Columbia University
Award #**1939192** | Award Period: 9/01/2019-2/28/2021

Abstract: The conference purpose is to stimulate a national conversation concerning the relationships between assessment, teaching and learning that include scholarly research and development of tests; members of city and state boards of education; officials from states and major school systems; policymakers; and representatives of teachers' associations and parents' associations. This conference aims to attract these important professionals has important co-sponsors like the Urban Institute. This national conference flows from the work of the Gordon Commission on the Future of Assessment for Education that addressed the advancement of achievement in STEM disciplines (PRE-K - 12) for students who are underrepresented among high achieving students. This issue of advancement of underrepresented high achieving students has received little concentrated effort and a conference would help in providing greater understanding of this special concern, which includes a student in poverty in complexed family structures.

Developing a Suite of Standards-based Instructionally Supportive Tools for Middle School Computer Science

Principal Investigator: Satabdi Basu, SRI International
Award #**2010591** | Award Period: 5/15/2020-4/30/2024

Abstract: This project will develop a set of educational resources, assessment tools and teacher professional development (PD) activities to support teachers in developing knowledge of CS content and standards. Improved CS instruction that is responsive to the needs and challenges of the student population is particularly critical in school districts with a large population of students who are typically underserved and under-represented in computer science. The project, a partnership between SRI International, the Milwaukee Public School District, and the San Francisco United Public School District, will provide professional development experiences tied to standards. Teachers will receive training via a combination of virtual webinars and face-to-face instruction. Teachers will have opportunities to evaluate their own teaching and measure their students' progress towards the standards.

Supporting Science Learning and Teaching in Middle School Classrooms through Automated Analysis of Students' Writing

Principal Investigator: Rebecca Passonneau, Pennsylvania State Univ. University Park
Award #**2010351** | Award Period: 8/01/2020-7/31/2024

Principal Investigator: Sadhana Puntambekar, University of Wisconsin-Madison
Award #**2010483** | Award Period: 8/01/2020-7/31/2024

Abstract: This project will develop a novel, automated technology to provide middle-school students and their teachers with real-time feedback about students' written explanations of physics phenomena. Through the four-year project, an iterative development process will include the design of the system and testing of two iterations of the system; research of student responses across the progression of roller coaster design and written assignments; and use of some validated and custom instruments to assess student understanding of key forces and assessment of student abilities to use data to evaluate claims. Classroom studies will use video data and researcher field notes to help understand how teachers facilitated the use of the wise-crowd system. Research will culminate in testing of the final version of the wise crowd system. Using a quasi-experimental design, classes will be randomly assigned to the treatment or comparison conditions. Findings will advance knowledge in the field about the best ways to integrate content assessment and feedback from the automated system with classroom and individual support from teachers to optimize learning for students. Materials and results generated from the project will be broadly disseminated, resulting in significant impacts for researchers and practitioners.

Preparing Teachers to Design Tasks to Support, Engage, and Assess Science Learning in Rural Schools

Principal Investigator: William Penuel, University of Colorado at Boulder
Award #**2010086** | Award Period: 9/01/2020-8/31/2024

Abstract: Researchers at Colorado University Boulder and BSCS Science Learning will use design-based implementation research to collaboratively design an online course sequence that targets 5D assessment in science. The study will proceed in three phases: a rapid ethnographic study to assess the needs of teachers serving a variety of rural communities, a study of teachers' use of an online platform for their professional learning, and lastly an experimental study to research the effects of the online course on teacher and student outcomes. Researchers will recruit 10 teachers to take the on-line course for the professional development and collect data on participating teachers' implementation of the course ideas through classroom videotaping and surveys designed to capture their changing practices. In the third year of the project, researchers will conduct an impact study with 70 secondary science teachers taking the re-designed on-line course, and compare their outcomes with a "business-as-usual" condition.

AI-based Assessment in STEM Education Conference

Principal Investigator: Xiaoming Zhai, University of Georgia Research Foundation Inc
Award #**2138854** | Award Period: 8/01/2021-7/31/2023

Abstract: The conference is organized around four themes: (a) AI and Domain Specific Learning Theory; (b) AI and validity theory and assessment design principles; (c) AI and technology integration theory; and (d) AI and pedagogical theory focusing on assessment practices. It allows participants to share theoretical perspectives, empirical findings, as well as research experiences. It can also help identify challenges and future research directions to increase the broad use of AI-based assessments in science education. The conference will be open to other researchers, postdocs, and students via Zoom. It is expected that conference participants establish a network in this emergent area of science assessment. Another outcome of the conference, Applying AI in STEM Assessment, will be published as an edited volume by Harvard Education Press.

Collaborative Research: Developing and Evaluating Assessments of Problem-Solving in Computer Adaptive Testing Environments

Principal Investigator: Jonathan Bostic, Bowling Green State University
Award #**2100988** | Award Period: 8/01/2021-7/31/2026

Abstract: This project expands the scope of the problem-solving measures use and score interpretation. The project work advances mathematical problem-solving assessments into computer adaptive testing. Computer adaptive testing allows for more precise and efficient targeting of student ability compared to static tests. Few measures designed to assess students' mathematical problem-solving ability use this technology. Shorter tests require less in-class time for assessment than current paper-pencil problem-solving measures and increase classroom instruction time. The computer-adaptive problem-solving measures have sufficient reliability and strong validity evidence, and may limit test-taker fatigue. Finally, the project will benchmark current grades 6-8 instruments using an objective standard-setting method, which allows for improved score interpretations with content-related feedback. Immediate results of student- and class-level reports will be produced through the computer adaptive testing system allowing for teachers to modify instruction to improve students' learning.

Empowering Teachers to See and Support Student Use of Crosscutting Concepts in the Life Sciences

Principal Investigator: Chad Gotch, Washington State University
Award #**2100822** | Award Period: 8/01/2021-7/31/2024

Abstract: The project focuses on the development of formative assessment tools that highlight assets of students' use of crosscutting concepts (CCCs) while engaged in science and engineering practices in grades 9-12 Life Sciences. In response to the calls set forth by the Framework for K–12 Science Education and Next Generation Science Standards (NGSS), the field has most successfully researched and developed assessment tools for disciplinary core ideas and the

science and engineering practices. The CCCs, which serve as the connective links across science domains, however, remain more abstractly addressed. Presently, science educators have little guidance for what student use of CCCs looks like or how to assess and nurture such use. This project, with its explicit attention to the CCCs, advances true three-dimensional scientific understanding in both research and the classroom.

Supporting Instructional Decision Making: The Potential of An Automatically Scored Three-dimensional Assessment System

Principal Investigator: Christopher Harris, WestEd
Award #**2101112** | Award Period: 9/01/2021-8/31/2025

Principal Investigator: Joseph Krajcik, Michigan State University
Award #**2100964** | Award Period: 9/01/2021-8/31/2025

Principal Investigator: Xiaoming Zhai, University of Georgia Research Foundation Inc.
Award #**2101104** | Award Period: 9/01/2021-8/31/2025

Principal Investigator: Yue Yin, University of Illinois at Chicago
Award #**2101166** | Award Period: 9/01/2021-8/31/2025

Abstract: This project will study the utility of a machine learning-based assessment system for supporting middle school science teachers in making instructional decisions based on automatically generated student reports (AutoRs). The assessments target three-dimensional (3D) science learning by requiring students to integrate scientific practices, crosscutting concepts, and disciplinary core ideas to make sense of phenomena or solve complex problems. Led by collaborators from University of Georgia, Michigan State University, University of Illinois at Chicago, and WestEd, the project team will develop computer scoring algorithms, a suite of AutoRs, and an array of pedagogical content knowledge supports (PCKSs). These products will assist middle school science teachers in the use of 3D assessments, making informative instructional changes, and improve students' 3D learning. The project will generate knowledge about teachers' uses of 3D assessments and examine the potential of automatically scored 3D assessments.

Developing Science Assessments for Language Diversity in Early Elementary Classrooms

Principal Investigator: Daisy Rutstein, SRI International
Award #**2201051** | Award Period: 8/01/2022-7/31/2026

Abstract: This project will design instructional assessment materials by using an innovative and unique design approach that brings together the coherent and systematic design elements of evidence-centered design, an equity and inclusion framework for the design of science materials, and inclusive design principles for language-diverse learners. Using this three-pronged approach, this project will develop a suite of NGSS aligned formative assessment tasks for first-grade science and a set of instructional materials to support teachers as they administer the formative assessments to students with diverse language skills and capacities. These resources will be of high utility for classroom teachers in the early elementary grades and offer an important step towards meeting the need for usable NGSS aligned assessments for the early grades. The design approach will provide assessment developers with a rigorously designed and tested model for developing their own NGSS-aligned assessments which will support the study of science in early elementary classrooms and promote further development of new curriculum materials, and support the evaluation of existing curriculum materials.

Completing the Development of the Electronic Test of Early Numeracy (e-TEN)

Principal Investigator: Arthur Baroody, University of Illinois at Urbana-Champaign
Award #**2201039** | Award Period: 9/01/2022-8/31/2026

Abstract: This project continues and completes the development and refinement of an electronic Test of Early Numeracy (e-TEN) in English and Spanish, focused on number and operations for young learners. The assessment incorporates a learning trajectory approach that describes students' development of the understanding of number. The electronic assessment allows for the test to adapt to students' responses and incorporate games increasing children's engagement with the tasks. These features take advantage of the electronic format. The achievement test is designed to be efficient, user-friendly, affordable, and accessible for a variety of learning environments and a broad age range of students (3 to 8 years old). The goal of the assessment design is to create a measure that is more accurate, more accessible to a wider range of children, and easier to administer than existing measures. The proposed activities also include establishing 10 separate but equated computer adaptive tests to produce a complete commonly scaled construct for ages 3-8, capturing change in numeracy ability across 10 six-month intervals. These efforts will be validated in both English and Spanish with national norms.

Evaluating Effects of Automatic Feedback Aligned to a Learning Progression to Promote Knowledge-In-Use

Principal Investigator: Kevin Haudek, Michigan State University
Award #**2200757** | Award Period: 9/01/2022-8/31/2026

Abstract: This project examines the effect of an assessment system that automatically generates feedback based on students' open-ended assessment responses in chemistry and physics consistent with a previously developed learning progression that describes the successively more complex understandings students can develop about electrical interactions. The researchers will design and test an automated assessment scoring system using machine learning. The scoring system will provide individualized feedback to students and class summaries to their teachers. This could then serve as a formative assessment to match an existing high school physical science curriculum designed to meet performance expectations in the Next Generation Science Standards. The project will then examine whether the automatic feedback supports students' learning outcomes and their development with respect to the learning progression on electrical interactions. The project promotes students' knowledge of science by engaging them in scientific practices, like modeling, with key disciplinary ideas and using crosscutting concepts to make sense of compelling phenomena and by providing real-time feedback to students. Deepening students' knowledge of science requires that they have opportunities to solve ill-structured, complex problems and to create models of real-world phenomena. This project serves the national interest by examining how to assess and support students in responding to such problems in science through timely and productive feedback about their performances.

Community for Advancing Discovery Research in Education

CADRE is a network for STEM education researchers funded by the National Science Foundation's Discovery Research PreK-12 (DRK-12) program. Through in-person meetings, a website, common interest groups, newsletters, and more, CADRE connects these researchers who are endeavoring to improve education in science, technology, engineering, and mathematics in, and outside of, our schools.

CADRE helps DRK-12 researchers share their methods, findings, results, and products inside the research and development community and with the greater public so that we are:

- **Better informed** about the work that is being done,
- **Continually building** on what we have collectively learned,
- **Working with our schools, communities, and policy-makers** to make our findings and products accessible and usable, and
- **Progressively able to address new and more challenging issues—**including those issues that extend beyond the limits of what any singular research project can impact.

Together, we can make a larger impact on policy, research, and education.

Contact cadre@edc.org for more information.

EDC Education Development Center