Emergent Design Heuristics for Three-Dimensional Classroom Assessments that Promote Equity

Erin Marie Furtak, University of Colorado Boulder, erin.furtak@colorado.edu
Hosun Kang, University of California Irvine, hosunk@uci.edu
James Pellegrino, University of Illinois Chicago, pellegjw@uic.edu
Christopher Harris, WestEd, christopher.harris@wested.org
Joseph Krajcik, Michigan State University, krajcik@msu.edu
Deb Morrison, University of Washington, eddeb@uw.edu
Philip Bell, University of Washington, pbell@uw.edu
Heena Lakhani, University of Washington, hlakhani@uw.edu
Enrique Suárez, University of Massachusetts Amherst, easuarez@umass.edu
Jason Buell, University of Colorado Boulder, jason.buell@colorado.edu
Jasmine Nation, University of California Irvine, nationj@uci.edu
Kate Henson, University of Colorado Boulder, kate.henson@colorado.edu
Caitlin Fine, University of Colorado Boulder, caitlin.fine@colorado.edu
Paul Tschida, Tustin High School, ptschida@tustin.k12.ca.us
Lindsay Fay, Tustin High School, lfay@tustin.k12.ca.us
Quentin Biddy, University of Colorado Boulder, quentin.biddy@colorado.edu
William Penuel, University of Colorado Boulder, william.penuel@colorado.edu
Kerri Wingert, University of Colorado Boulder, kerri.wingert@colorado.edu

Abstract: In 2014, the National Research Council posited five criteria for assessments that engage students in scientific practices, crosscutting concepts, and disciplinary core ideas (also known as "three-dimensional assessment"). Multiple efforts have been conducted to design and study three-dimensional assessments in K-12 science classrooms; yet these projects have employed different approaches while generating different guidelines and tools. This symposium brings together members of five research teams to reflect on the NRC criteria using examples of three-dimensional assessments. Through a collective analysis of these teams' current design approaches, we identified six emergent criteria for three-dimensional assessment design that promote equity: (1) assessments are based on a relevant phenomenon or scenario, (2) explicitly attend to language, (3) include scaffolds to make expectations explicit for students, (4) attend to the identities of learners, (5) engage and support student sensemaking, and (6) are accompanied by tools and routines to support teacher design and enactment.

Introduction

Ever since release of the *Framework for K-12 Science Education* (National Research Council [NRC], 2012), educators, science leaders, and researchers have grappled with its implications for the design and enactment of classroom assessments; that is, the assessments engaged in by teachers and students during the course of instruction. The NRC articulated a three-dimensional vision of science learning, arguing that students should engage in science and engineering practices, and apply crosscutting concepts as students learn disciplinary core ideas. Additionally, the *Framework* emphasizes that student opportunities around three-dimensional learning should promote equity by broadening the participation of students historically marginalized in science.

As the majority of US states have now adopted science standards based on the *Framework*, designing and enacting three-dimensional classroom assessments that promote equity are a responsibility for the field. The NRC (2014) provided early guidance for the design of classroom assessments based on reviews of literature in science education and assessment design, asserting that they should (a) include multicomponent tasks, (b) be based on a progression of learning, (c) provide learners with multiple opportunities to demonstrate their learning, (d) be related to bundles of performance expectations [PE's] from the Next Generation Science Standards [NGSS], and (e) be designed to promote equity and justice in science classrooms.

This symposium reflects on the ways in which these original design criteria for three-dimensional classroom assessment have been taken up by five research teams deeply engaged in science assessment design at different scales. Synthesizing across our projects, we identify a set of emergent design criteria, as well as tensions in promoting equity. Making these design heuristics explicit is a crucial step as we seek broader participation for students from historically marginalized populations in STEM.

Overview of individual presentations

Each of the teams will briefly provide context for their three-dimensional assessment design projects, and then will provide rich and specific examples of how the criteria listed in Table 1 are embodied in their task designs.

Team 1: Designing and using classroom-based assessments to assess growth towards knowledge-in-use and the NGSS

We have designed, tested and refined a procedure for developing classroom-based assessments that align with the NGSS and promote students' knowledge-in-use to make sense of phenomena. The assessments provide information to teachers and learners on using knowledge that aligns with learning goals that integrate disciplinary core ideas, crosscutting concepts and scientific practices. Our design process is systematic, scalable and equity-focused for designing assessment tasks that measure student three-dimensional science proficiency.

Design heuristics

How to design NGSS-aligned assessments presents a challenge for assessment designers and science educators. We developed a modified Evidence-Centered Design approach ([ECD]; e.g., Almond, Steinberg, & Mislevy, 2002) that supports the development of tasks using the three dimensions of science proficiency. Our approach, like ECD, emphasizes the evidentiary base for specifying coherent, logical relationships among (1) learning goals that comprise the constructs to be measured (i.e., the claims we want to make about what students know and can do); (2) evidence in the form of observations, behaviors, or performances that should show the target constructs; and (3) features of tasks or situations that should elicit those behaviors or performances. We tackled developing assessment tasks in which learners need to make use of the three dimensions (Harris, Krajcik, Pellegrino & DeBarger, 2019). The tasks are also technology-enhanced (e.g., use of simulation, modeling software, video) and many use non-textual representations to elicit student responses (e.g., through drawing or modeling).

Using the modified ECD approach, we start by systematically unpacking the 3-dimensional performance expectations [PEs] that are used in the NGSS. Because PEs have a large grain size, we use the unpacking to develop smaller, more manageable learning goals that remain three-dimensional. We refer to these intermediate three-dimensional learning goals as *learning performances,* or knowledge-in-use statements that guide the development of assessment tasks and rubrics for measuring students' progress towards achieving three-dimensional aspects of the PEs. Our overall design process involves three distinct phases that are iterative and recursive - Domain Analysis, Domain Modeling & Task Development (see Harris et al, 2019).

Attention to equity

We consider equity and inclusion throughout our design process (Alozie et al., 2018). We have unpacked multiple performance expectations from physical and life science, constructed learning performances, and used the learning performances to create over 100 assessment tasks for use in middle school classrooms. These tasks can be found at the Next Generation Science Assessment website. To date, over 44,000 users have visited our assessment portal. Our systematic design process helps to ensure that we have valid assessment tasks that build towards learners meeting the performance expectations of the NGSS. We have also used a series of studies to ensure validity and reliability of the tasks including cognitive lab studies to provide data on task comprehensibility and issues of equity and construct-irrelevant variance. Performance studies have provided data on item features (e.g., such as item difficulty) that affect student performance and on the utility of our rubric design.

Team 2: Broadening systems of formative assessment to inform cognitive and cultural aspects of instruction

We synthesize learnings from several multi-year collaborative design projects in a variety of educational organizations such as school districts, state-level educational networks, and cross-state science education collaborations in the US. Our work to expand what constitutes formative assessment arises out of sustained engagements in these contexts to support educators in understanding and operationalizing equity in the context of implementation efforts around the *Framework* (NRC, 2012). In each of these partnerships we engaged in collaborative design to address challenges faced by educators that offered opportunities in practice to improve equity. Several shared challenges among our partners emerged related to formative assessment. For example, educators were eager for further guidance on how to make sense of students' everyday ideas about scientific phenomena, creating opportunities to grow their understandings of the intellectual treasures (knowledge and practices) students bring into the learning context. Additionally, educators were keen to identify and understand the varied ways their students struggled to engage with learning activities, which created an opportunity to improve student learning. Each of these challenges and opportunities could be addressed through co-developing

tools for assessment that support educators' interpretive power (Rosebery, Warren, & Tucker-Raymond, 2016).

Design heuristics

We have focused on understanding what constitutes a system of formative assessment in practice. Historically, formative assessment has measured what people are cognitively learning. Recently, this has expanded to include both conceptual knowledge and epistemic practices (Ford & Forman, 2006) as a result of the current focus within science education for complex multi-dimensional formative assessment activity. Our first design principle is grounded in a system of assessment perspectives and proposes that formative assessment activity must *identify supports for educators to interpret and use students' responses about what they know and can do related to such assessment tasks*. Our work also builds on improvement science's efforts to understand the nature of learning activity through practical measures (Yeager et al., 2013). As science and science learning are cultural activities situated within socio-historical contexts, we expect that educators and students will experience learning activity differently. Practical measures provide educators with insights to understand student learning experiences and adjust instruction. Thus, our second design principle is to *support educators to interpret and adjust instruction to improve equitable participation in learning activity by responding to learner interest and identities* (Tzou & Bell, 2010). A third principle is that educators *learn through iteration* (McDonald, et al., 2014), and as such *assessment must also be constantly modified and contextualized in practice* (Ruiz-Primo & Furtak, 2007) ensuring that assessment, like instruction, is locally relevant.

We present two examples of how we have operationalized these design principles in practice to expand a system of formative assessment within science education. First, we have developed supports around the construction of equitable multi-dimensional formative assessments that are designed to assess scientific concept and practice learning goals, and are designed to ensure students can equitably engage in the assessment activity. These tools have been used with thousands of teacher educators and teachers in a variety of learning contexts to produce localized multi-dimensional formative assessment tasks. Educators report how understanding the ways assessments can be inequitable has impacted their pedagogy; the seeds of these changes were evident in the assessments participating educators modified and/or created during the workshops. Second, we have developed resources to support educators' identification of facets of students' ideas (Minstrell, 1992) and practices that we then use to support next steps of instruction. In one of our partnerships, a group of middle school science teachers spent a year reviewing and revising the full suite of in-classroom formative assessments they used; a comparison between the beginning and final iterations of these assessments shows an intentionality in the design to ensure that teachers gained better insight into student thinking and doing in order to inform the next lessons.

Attending to equity

We found that there was still a need to expand interpretive power around the varied experiences students have in learning contexts resulting from socio-historic issues of inequity that may be present and yet not visible to educators (Rosebery et al., 2015). To this end we have developed, tested and refined practical measures at multiple scales across the U.S. to provide guidance to formatively assess *how* students are experiencing learning. The items in such assessments include Likert scale prompts that ask about the frequency of specific scientific practices in which students are engaged, or the degree of agreement around how instructional activity connected with a student's life. When collected frequently, such measures help to identify areas where particular students are engaged or disengaged from learning, informing possible areas of focus for instructional reform to improve equity of participation through constant modification and contextualization of learning activity. Collectively, these types of formative assessment form a system of assessment that can support efforts towards equitable science education.

Team 3: Tools for co-design of formative assessments of a crosscutting concept

The *Framework* vision of three-dimensional learning includes crosscutting concepts, or ideas that span across the disciplines such as patterns, scale, and structure and function (NRC, 2012). The crosscutting concepts reflect the kinds of questions that scientists might ask when presented with a novel scenario and form an important pathway through which students might make sense in the context of classroom assessment. Since 2015, our partnership between university-based researchers, teachers and curriculum coordinators in a large school district has designed and facilitated school-based, content-specific professional learning community (PLC) meetings focused on the design, enactment, and reflection upon three-dimensional formative and summative assessments aligned with the crosscutting concept of energy and matter flows, cycles and conservation. Working from a sociocultural perspective on learning, we have examined the ways in which a system of tools can organize routines for teacher participation in co-design that in turn supports shifts in classroom practice. Our design work is centered around a learning progression for modeling energy and matter flows within systems (Buell et al., 2019).

Design heuristics

The partnership has designed sets of tools and routines to coordinate teacher co-design of formative assessments across school sites. Teachers first explore student thinking in the domain of energy by interpreting student work with the support of a learning progression. Then, teachers work with their existing curriculum materials to design an embedded formative assessment task. They are supported in this design with the use of a formative assessment design checklist, developed on the basis of NGSS resources and prior research (e.g. Achieve, Inc., 2017; Kang et al., 2014). Teachers then determine the ways in which that task will support student participation structures in the classroom. After teachers enact the task with students, they examine student work with the learning progression, determining what students know and how they might be supported in subsequent learning. Throughout this process, we have collected qualitative data such as audiorecordings of PLC meetings, facilitator fieldnotes, teacher-designed formative assessment tasks, and student work. Our ongoing analyses of these sources of data have allowed us to iteratively refine our facilitation guides, our checklist, and our learning progression.

Rather than working with bundled PE's, our approach bases assessment design on a single performance expectation that includes modeling energy both as a crosscutting concept and a disciplinary core idea, and then designs specific questions for that task on the basis of PE evidence statements in the NGSS. The co-designed tasks that result consist of multiple open-ended components, including space for students to draw and label models, use their model to explain a phenomenon, and respond to other related questions. In addition, each question in the co-designed formative assessment tasks relates to a contextualized phenomenon. We have developed these tasks in relation to evidence outcomes related to individual performance expectations, instead of *bundles* of standards.

Attending to equity

Students in our partner district speak multiple languages and dialects, so formative assessment tasks must create space for learners from heterogeneous linguistic backgrounds to show what they know. We have designed a checklist to support science assessment task design to validate and create space for emergent bilingual learners to use multiple linguistic resources for sense-making and communication (Fine & Furtak, in press). These tasks are linguistically modified and invite students to use all of their linguistic resources as they process and respond to tasks. While we have used versions of this checklist in some settings with teachers and district-level coordinators, we are now examining how this checklist can inform the design of multiple forms of classroom assessment.

Team 4: Co-designing three-dimensional classroom assessments to improve opportunities to learn

A team of high school teachers, science education researchers, and university-based scientists formed a research-practice partnership to improve a local STEM learning ecology. The goal was to promote complex thinking in STEM for youth from non-dominant communities as responsible citizens while addressing the NGSS. Grounded in sociocultural and critical perspectives on learning (Greeno, 2006; Ladson-Billings, 1995), and building upon the perspective of assessment *for* learning (Black, Harrison, Lee, Marshall, & Wiliam, 2004), three premises guide our work of designing and enacting three-dimensional assessment: (a) assessment creates a particular form of opportunity for students to show their proficiency, inherently privileging a particular way of thinking and talking over others; (b) assessment can and should support meaningful science learning of a wide range of learners; and (c) improving assessment can improve learning and also reduce inequity and injustice at public schools.

Design heuristics

In early 2018, the team co-designed a set of curriculum and assessments in the unit of the rate of chemical reaction and equilibrium for high school chemistry. The unit storyline included: (a) a focal phenomenon and the essential question (i.e., why does [our town] have more severe and frequent wildfires now than 100 years ago?), (b) initial and final written assessment and a rubric, and (c) a sequence of key learning activities in the unit. The participating teachers implemented the curriculum in 10th/11th grade chemistry classrooms for 5.5 weeks in spring of 2018. The school served a large number of students from Latinx, immigrant, and low-income family backgrounds. The curriculum and assessments were revised and enacted once again in spring 2019. The final written assessment task was to construct evidence-based explanations. In addition, teachers enacted a non-traditional form of assessment (i.e., digital storytelling—create a PSA answering, "What is an action you can take to lower the frequency or severity of wildfires near [our town]?"). The non-traditional assessment created expanded opportunities for students to show what they could do by leveraging their funds of knowledge and to engage in their identity work (Calabrese-Barton, Kang, Tan, O'Neil & Bautista-Guerra, 2012). The following questions guide our inquiry: 1) How did students respond to the three-dimensional written assessment tasks? Who did well and who didn't? Were there any differences between different groups of students in their performances? 2) How did improving the design of three-dimensional written assessment tasks affect students' engagement in scientific

sense-making over two years, if at all? 3) How did different groups of students respond to two different designs of three-dimensional assessments (original vs. revised)? Through the two years of partnership activities for improvement, five assessment design heuristics emerged: (a) selecting a locally contextualized phenomenon or problem that matters to students or communities, (b) framing a question in a way of highlighting key observations or patterns (i.e., the subject of sense-making), (c) creating multiple forms of opportunities for students to show their ideas and identities, (d) embedding various built-in scaffolds that address linguistic, intellectual, and relational challenges in engaging in disciplinary discourses and practices, and (e) making 'the rules of game' (i.e., disciplinary way of thinking and talking) explicit by design.

Attending to equity

This study reveals two tensions in promoting equity. One is assessing an individual student's sense-making in a classroom community where collaborative and collective knowledge building is highly valued and promoted. The other tension is reducing the complexity of real-world phenomena or problems in order to make the work of sense-making accessible to students while keeping its scientific integrity.

Team 5: From three to five dimensions: Design heuristics for science assessments that elicit interest and identity

Research points to the importance of interest and identity for promoting learning, particularly for students from nondominant communities (National Academies of Sciences Engineering and Medicine [NASEM], 2018; NRC, 2012). Culture-based pedagogies can support equity by helping students make connections between science content and their interests and identities (e.g., Bang & Medin, 2010; Tzou & Bell, 2010). Assessment can support such pedagogies, but doing so requires an expansive conception of assessment (Penuel & Shepard, 2016) that go beyond gauging mastery of standards to elicit information about students to connect curriculum more meaningfully to students' lives and facilitate relationship building (Penuel & Watkins, 2019).

Design heuristics

We conceptualize the use of assessment adaptation as a cycle that involves planning, enacting, and reflecting on assessments (Furtak, Circi, & Heredia, 2018). Our design begins with the premise that equitable formative assessment must be grounded in a coherent theory of learning and supported by curriculum aligned with that theory (Shepard, Penuel, & Pellegrino, 2018). This curriculum supports students' agency, positioning students as constructors and critiquers of science knowledge, and is anchored in phenomena selected using evidence of student interests. The classroom routines support students in making connections across and between lessons and their everyday lives. The assessments, in the form of a Student Electronic Exit Ticket (SEET), are intended to be practical to administer and use. We design these "practical measures" to be lesson-focused and include a set of prompts anchored in the day's investigative phenomenon, pairing two dimensions of science learning (core idea-practice, core idea-crosscutting concept, crosscutting concept-practice). Tasks include questions about students' perception of their contributions to and the relevance and coherence of the lesson. Teachers choose a classroom routine to adapt that addresses an equity issue identified from their practice.

Attending to equity

Using principles of improvement science, we engage teachers in Plan-Do-Study-Act cycles where they (1) collect data about equity of participation related to race, gender, and home language as part of an electronic "exit ticket," (2) analyze data and identify strategies to test in classrooms for improving equity of participation, and (3) analyze and use evidence related to improvement to adjust their teaching. We focus on equity of participation, in order to promote epistemic justice (Fricker, 2009) at the classroom level, to re-position students from nondominant groups as science knowledge builders in the classroom (Miller, Manz, Russ, Stroupe, & Berland, 2018).

Summary: Emergent design heuristics

Each research team summarized their work with three-dimensional assessments with respect to: design approach, emerging principles, tools or guidelines used to design the assessment tasks/items/system, key task features, and when and how they attended to equity. The first two authors engaged in open and closed coding, examining each project's approach to designing 3D assessment with respect to these four dimensions. Through that process of coding, we identified six emergent heuristics, then shared with the other authors who provided feedback to further inform and refine the heuristics and their descriptions.

All five teams featured multicomponent tasks that provide students with multiple opportunities to show their learning. However, across teams, there were variations in terms of working with bundles of PE's (d) as compared to a smaller grain size - revealing a tension between what can be easily assessed and useful for teachers

and addressing the number of available standards to be assessed in a given grade band. In addition, some teams worked explicitly with progressions of learning while others based their designs on NGSS Performance Expectations, which are based on progressions of learning in the *Framework* (b). Finally, all projects described explicit attention to equity in some form (e), although there were notable differences in terms of when and how they attended to the equity in designing assessments. Taken together, a total of six new design heuristics emerged: (1) use of phenomena, (2) explicitly attending to language, (3) appropriate use of scaffolds to make explicit expectations for students, (4) explicit attention to identities of learners, (5) supporting student sense-making, and (6) accompanying by tools and routines to support teacher design/adaptation and enactment (Table 1).

Table 1. NRC Three-dimensional task design criteria and emergent design heuristics

Original NRC (2014) design criteria	Emergent design heuristics (2020)
a. Multicomponent tasks b. Based on progressions of learning c. Provide multiple opportunities for students to show their learning d. Based on bundles of performance expectations e. Attention to equity	1. Based on a relevant phenomenon or scenario 2. Explicit attention to language 3. Includes scaffolds to make expectations explicit for students (e.g. disciplinary ways of thinking and talking) 4. Explicit attention to identities of learners 5. Engages and support student sensemaking 6. Accompanied by tools and routines to support teacher design/adaptation and enactment

Based on relevant phenomena or scenarios

All five teams connected tasks to phenomena, or real-world scenarios that serve as the basis for assessment questions. These phenomena are intended to connect students' cultural backgrounds and experiences with the science they are learning in school (e.g. NASEM, 2019). Teams described phenomena as contextualized, rather than separate from everyday examples, and interesting to students. Some teams emphasized they should also be of local or global significance, and of concern to students and interest to the communities in which students live.

Explicit attention to language

The language used on assessments was also an area of focus for several teams. To increase accessibility of NGSS-aligned tasks, assessments must consider the heterogeneous linguistic practices and cultural resources of students as they show what they know (Mislevy & Durán, 2014) in addition to multimodality of assessment tasks. Three of the teams explicitly noted attention to the linguistic complexity of tasks, creating space for students to share their ideas using everyday language, and even allowing students to respond to task in languages other than English.

Inclusion of scaffolds to make expectations explicit for students

Several teams noted the norms and expectations built into task design. Showing disciplinary proficiency involves participating in communities of practice that privilege a particular way of thinking, doing, and talking (NRC, 2012). Building upon the argument of critical scholars who problematize the settled hierarchy as a fundamental challenge for achieving equity (Bang, Warren, Rosebery, & Medin, 2012), the teams identified the inclusion of scaffolds to make the 'rules of the game' are explicit by design for students from non-dominant communities.

Explicit attention to the identities of learners

Drawing upon sociocultural perspectives, three teams noted that students' opportunity to access and engage in disciplinary tasks is substantially affected by the connection established between the learner and the task. Creating multiple entry points for students from heterogeneous backgrounds in a particular classroom setting necessitates designers' attention to learners and their interests and concerns (Calabrese-Barton & Tan, 2010). However, there were nuanced differences across the teams in terms of how the identities of learners were considered in the process of designing assessment tasks. For example, two teams systematically collect and use the information about the identities primarily to select the phenomena. One team also highlighted creating space for identity work by designing assessment tasks that not only show what they know but also who they are.

Engage and support student sensemaking

A smaller number of teams identified sensemaking as a core element of three-dimensional assessment. These teams attend to the goal of science learning advocated by the NGSS, supporting students to actively make sense

of the world (Bang, Brown, Calabrese-Barton, Rosebery & Warren, 2017; NRC, 2012; Schwarz, Passmore, & Reiser, 2017). Teams identified engaging and supporting students' sensemaking as an important design heuristic to create contexts for tasks that create opportunities to assess students' developing knowledge and practice.

Accompanied by tools and routines to support teacher adaptation and enactment

Most of the groups integrated supports for teachers designing and enacting three-dimensional tasks. These tools and routines respond to research that has identified challenges teachers face when enacting classroom assessment in general (Sezen-Barrie & Kelly, 2017), and three-dimensional assessment in particular (Furtak, 2017). While one team designed tasks and made them available to teachers and students via an online system, four other teams engaged multiple sets of tools to provide for teachers in designing and using tasks themselves, facilitation guides for iterative cycles of design and enactment, task design checklists, and tools for interpreting student work.

Significance

As more learning environments adopt and implement the vision of the *Framework*, researchers and educators must account for the ways in which assessment can maintain and/or disrupt inequitable learning opportunities. We intend that our presentations can also raise questions about whose knowledge and expertise should be leveraged in designing three-dimensional classroom assessments to promote equity. Our intention is that, by identifying emergent design criteria, we will help move the field toward a consensus on what quality classroom assessments aligned with the *Framework* vision look like in order to inform subsequent design, implementation, and use of those assessments in classrooms. The papers in this session present systematic design processes, tools, examples, and detailed analyses of high-quality, three-dimensional assessments that elicit student learning consistent with the vision of the *Framework* and that align with the performance expectations of the NGSS.

References

Achieve, Inc., (2017). *Science Assessment Task Screening Tools.* https://www.nextgenscience.org/taskscreener

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1 (5).*

Alozie, N., Pennock, P.H., Madden, K., Zaidi, S., Harris, C., & Krajcik, J. (2018, March). *Designing and developing NGSS-aligned formative assessment tasks to promote equity*. Paper presented at the annual conference of National Association for Research in Science Teaching, Atlanta, GA.

Bang, M., Brown, B. A., Calabrese-Barton, A., Rosebery, A., & Warren, B. (2017). Toward more equitable learning in science: Expanding relationships among students, teachers, and science practices. In C. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 33-58). Washington, DC: NSTA.

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

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

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

Buell, J.Y., Briggs, D.C., Burkhardt, A., Chattergoon, R., Fine, C., Furtak, E.M., Henson, K., Mahr, B., & Tayne, K. (2019). A learning progression for modeling energy flows in systems. Boulder, CO: Center for Assessment, Design, Research and Evaluation (CADRE).

Calabrese-Barton, A., & Tan, E. (2010). Journal of the Learning We Be Burnin '! Agency , Identity , and Science Learning. *Journal of the Learning Sciences*, (January 2012), 37–41.

Fine, C. & Furtak, E.M. (in press). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education.*

Fricker, M. (2009). *Epistemic injustice: Power and the ethics of knowing*. New York: Oxford University Press.

Ford, M. J., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. *Review of Research in Education*, *30*(1), 1-32.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education*, virtual issue, 1–14. https://doi.org/10.1002/sce.21283

Furtak, E. M., Circi, R. K., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education, 31*(2), 143-156.

Greeno, J. G. (2006). Learning in activity. In K. R. Sawyer (Ed.), The cambridge handbook of the learning sciences (pp. 79-96). New York, NY: Cambridge University Press.
Harris, C., Krajcik, J., Pellegrino, J.W., & DeBarger, A. (2019). Designing Knowledge-In-Use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, Summer 2019, Vol. 38, No. 2,* 53–67.
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.
McDonald, M., Kazemi, E., Kelley-Petersen, M., Mikolasy, K., Thompson, J., Valencia, S. W., & Windschitl, M. (2014). Practice makes practice: Learning to teach in teacher education. *Peabody Journal of Education*, *89*(4), 500-515.
Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. *Research in physics learning: Theoretical Issues and Empirical Studies*, 110-128.
Miller, E. C., Manz, E., Russ, R. S., Stroupe, D., & Berland, L. (2018). Addressing the epistemic elephant in the room: Epistemic agency and the Next Generation Science Standards. *Journal of Research in Science Teaching, 55*(7), 1053–1075.
Mislevy, R. J., & Durán, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of common core and next generation science standards. *TESOL Quarterly*, *48*(3), 560–585.
National Academies of Sciences Engineering and Medicine. (2018). *How people learn II: Learners, cultures, and contexts*. Washington, DC: National Academies Press.
National Academies of Sciences, Engineering, and Medicine. (2019). *Science and engineering for grades 6-12: Investigation and design at the center.* Washington, D.C.: National Academies Press.
National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press.
National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.
NGSS Lead States. 2013. *Next Generation Science Standards: For states, by states.* Washington, DC: National Academies Press. *www.nextgenscience.org/ next-generation-science-standards.*
Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.),
Handbook of Research on Teaching (pp. 787-851). Washington, DC: AERA.
Penuel, W. R., & Watkins, D. A. (2019). Assessment to Promote Equity and Epistemic Justice: A Use-Case of a Research-Practice Partnership in Science Education. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 201-216.
Rosebery, A. S., Warren, B., & Tucker-Raymond, E. (2015). Developing interpretive power in science teaching. *Journal of Research in Science Teaching*, *53*(10), 1571-1600.
Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring Teachers' Informal Formative Assessment Practices and Students' Understanding in the Context of Scientific Inquiry. *Journal of Research in Science Teaching*, *44*(1), 57–84.
Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. Arlington, VA: NSTA Press.
Sezen-Barrie, A., & Kelly, G. J. (2017). From the teacher's eyes: facilitating teachers noticings on informal formative assessments (IFAs) and exploring the challenges to effective implementation. *International Journal of Science Education*, *39*(2), 181–212.
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.
Tzou, C., & Bell, P. (2010). Micros and Me: Leveraging home and community practices in formal science instruction. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (ICLS 2010) - Volume 1, Full Papers. (pp. 1127-1134). Chicago IL: International Society of the Learning Sciences.
Yeager, D., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement.* Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (Grant Nos. 1561751, 1561300, 1238253, 1316874, 1316903, 1316908, & 1903103); the Gordon and Betty Moore Foundation (4482); the Spencer Foundation (1556184 and 10011767), the Hewlett Foundation (1556127), and the Chan-Zuckerberg Initiative (2018-194933).