Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
xxxx 2019, Vol. 0, No. 0, pp. 1–15

Designing Knowledge-In-Use Assessments to Promote Deeper Learning

Christopher J. Harris, *WestEd STEM Program,* Joseph S. Krajcik, *CREATE for STEM Institute, Michigan State University,* James W. Pellegrino, *Learning Sciences Research Institute, University of Illinois at Chicago,* and Angela Haydel DeBarger, *The William and Flora Hewlett Foundation*

Contemporary views on learning highlight that deep learning occurs not simply by accumulating knowledge, but by using and applying knowledge as one engages in disciplinary activity. Increasingly, those concerned with education policy and practice are shifting priorities toward supporting deeper learning by emphasizing the importance of students' ability to apply knowledge in subject areas. Designers of student assessments are following suit and are taking on the challenge of creating a new generation of assessments. We present a principled approach for designing classroom-based assessments that not only assess deeper learning, but also provide teachers with critical information about how students are progressing toward achieving ambitious new learning goals. Our approach follows the evidentiary reasoning of evidence-centered design and builds on research about the important role of knowledge-in-use to support student learning. We illustrate our approach in the context of creating tasks that assess students' science proficiency as reflected in the Next Generation Science Standards that are gaining prominence in the United States.

Keywords: assessment design, deeper learning, science assessment

There has been an increasing discussion of ideas such as "deeper learning" and the development of 21st-century skills (e.g., Bellanca, 2014; Pellegrino & Hilton, 2012). The European Commission's *Rethinking Education* (2012) reform effort emphasizes the need to promote "transversal" skills in education, such as critical thinking and problem solving. Such 21st-century skills are deemed necessary to prepare a global workforce to succeed in a new information-driven economy. As discussed in the U. S. National Research Council Report *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st-century* (Pellegrino & Hilton, 2012), "Calls for such 21st-century skills as innovation, creativity, and creative problem solving can also be seen as calls for deeper learning—helping students develop transferable knowledge that can be applied to solve new problems or respond effectively to new situations." (p. 70).

The conceptual underpinnings for ideas such as "deeper learning" and "21st-century skills" are not new and are rooted in the desire to promote the development of knowledge and skills that transfer. As such, they draw from a rich legacy of work in the developmental, cognitive, and learning sciences over the past 40+ years, much of which is captured in synthesis reports like *How People Learn: Brain, Mind, Experience and School* (Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000) and *Education for Life and Work* (Pellegrino & Hilton, 2012). According to the latter report, deeper learning can be understood as the process through which a person becomes capable of taking what was learned in one situation and applying it to new situations—in other words, learning for transfer. Through deeper learning, individuals acquire expertise in a discipline or subject area that goes beyond the rote memorization of facts or procedures; they understand when, how, and why to apply what they have learned. They recognize when a new problem or situation is related to what they have previously learned, and they can apply their knowledge and skills to solve them. Deeper learning occurs not simply by accumulating knowledge, but by using and applying knowledge as one engages in disciplinary practices (Pellegrino & Hilton, 2012).

The need for integrated disciplinary knowledge and proficiency has gained prominence in the United States where new standards, notably the Common Core State Standards (CCSS; CCSS Initiative, 2010a, 2010b) addressing English language arts and mathematics, and the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) addressing science and engineering design, include

Christopher J. Harris, WestEd STEM Program, Redwood City, CA; christopher.harris@wested.com. Joseph S. Krajcik, CREATE for STEM Institute, Michigan State University, East Lansing, MI; krajcik@msu.edu. James W. Pellegrino, Learning Sciences Research Institute, University of Illinois at Chicago, Chicago, IL; pellegjw@uic.edu. Angela Haydel DeBarger, The William and Flora Hewlett Foundation, Menlo Park, CA; adebarger@hewlett.org.

goals that call for students to actively participate in the authentic practices of the disciplines. For example, both the U.S. mathematics and science standards include an emphasis on students using and applying knowledge in the context of disciplinary practices—that is, the actual everyday ways of knowing and doing that mathematicians and scientists employ in their respective fields. The basic premise 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, or make sense of phenomena. This knowledge-in-use perspective is garnering attention internationally in countries such as Finland, Germany and Singapore (e.g., Finnish National Board of Education, 2015; Kulgemeyer & Schecker, 2014) and has recently been emphasized in the domain frameworks for large-scale assessments, such as the Programme for International Student Assessment (PISA; OECD, 2016).

As noted above, the emphasis on knowledge-in-use reflects an increased awareness by educators, policymakers, and business leaders of the proficiencies required by people to participate as global citizens in the 21st century. It is widely recognized, for example, that to address the current and emerging challenges facing our world, our education systems will need to prepare a new generation of students who are able to use and apply knowledge in novel ways. This idea also presents a new way of thinking about what it means to be proficient within a given domain. In prior standards, proficiency was typically measured as acquisition of core content knowledge. In the new reform documents, it is not what you know, but how you use and apply what you know using disciplinary reasoning practices that demonstrate proficiency. From this perspective, knowledge-in-use is viewed as a means for promoting deeper learning, showcasing that learning, and preparing for future learning (Pellegrino & Hilton, 2012). It is a process for building and refining knowledge, solving problems, and investigating the natural world, and also a performance that represents deeper learning.

While there is tremendous interest and excitement for such new directions for education, it is not well known how to measure this complex, multidimensional learning; knowledge-in-use poses a formidable challenge to assessment design and validation. Unless students experience expectations reflecting deeper levels of learning, teachers will have limited information to support their students' multidimensional learning, and students will have little reason to develop the knowledge and skills that contemporary standards are meant to foster. Assessments communicate to multiple stakeholders what students are truly expected to know and be able to do. For assessment to be used formatively to support ongoing teaching and learning (e.g., Black & Wiliam, 1998), results must be informative relative to the domain of instruction and possible trajectories of student learning. How do we create assessments that require students to demonstrate rigorous and authentic evidence of these new learning goals? How do we measure students' progress during instruction and over time toward attaining them? How do we create assessments for classroom use that can help teachers come to know whether instructional experiences are building students' proficiency in using and applying knowledge through disciplinary practices and making real progress in their learning? To create valid, instructionally informative assessments, we must ensure that they were developed using a principled design approach that is transparent and documents intended claims of student proficiency. The assessment design framework must support the development of tasks that integrate two or more dimensions of knowledge and skill and to do so in ways that require and enable students to demonstrate capacity to make integrative use of core disciplinary ideas in the context of various practices. The resultant student responses should constitute explanations of phenomena and/or the design of solutions to problems (Pellegrino, Wilson, Koenig, & Beatty, 2014).

There have been noteworthy prior attempts to design assessments that reflect deeper engagement with disciplinary knowledge and cognitive processes, much of which is represented in efforts to develop performance assessments involving tasks of varying length and complexity. Such tasks have been incorporated into classroom assessment activities as well as state testing programs such as those of California, Vermont, Kentucky, and Maryland. The success of efforts to develop performance assessments has been mixed for various reasons. One problem has been the failure to specify in detail the cognitive underpinnings of the knowledge and skills to be assessed. A second problem has been the corresponding lack of a principled design process to guide the task development and validation processes (see e.g., Pellegrino, Chudowsky, & Glaser, 2001). Such challenges continue to persist with respect to performance assessment for contemporary disciplinary standards (see e.g., Davey et al., 2015; Tucker, 2015).

In this article, we present such a principled approach for designing classroom-based assessments that provide teachers with meaningful and actionable information about students' progress toward knowledge-in-use learning goals. Meaningful assessments offer experiences that provide "genuine, extended attention to the substance of student reasoning" and promote "awareness of how students are engaging in disciplinary practices" (Coffey, Hammer, Levin, & Grant, 2011, p. 1122). As actionable tools, we intend for information from assessments to be used by teachers to shape instruction in ways that address students' needs, interests, and experiences (Ruiz-Primo & Furtak, 2007). In addition to the tasks themselves, guidelines for how to make sense of students' ideas and actions in relation to learning goals are considered as part of the assessment design. In this way, we are aiming for a comprehensive assessment strategy to help teachers move students toward demonstrating these ambitious new learning goals.

Our approach draws from evidence-centered design (ECD) (Mislevy & Haertel, 2006), which has gained widespread attention as a comprehensive approach for principled assessment design and validation. ECD emphasizes the evidentiary base for specifying coherent, logical relationships among the (a) learning goals that comprise the constructs to be measured (i.e., the claims articulating what students know and can do); (b) evidence in the form of observations, behaviors, or performances that should reveal the target constructs; and (c) features of tasks or situations that should elicit those behaviors or performances. As such, ECD provides a framework for analyzing content for assessment design that can be used to specify the essential and assessable components of knowledge-in-use learning goals.

We illustrate our approach in the context of designing classroom-based assessment tasks that reflect the new vision for science proficiency presented in the National Research

Council's (NRC) *Framework for K-12 Science Education* (NRC, 2012) and instantiated in the knowledge-in-use learning goals of the NGSS. Since the 2013 release of the new standards, 19 states and the District of Columbia, representing more than 35% of the U.S. student population, have adopted them outright and more than 20 states have developed similar types of standards based on them. Furthermore, science leaders across the country have embraced the vision of the *Framework* and NGSS (e.g., National Science Teachers Association, 2016). States that have signed on to the NGSS, as well as curriculum developers, professional development experts, and classroom teachers need clear guidance on how to assess the knowledge and skills associated with the standards. As science teachers are called upon to shift their instruction to address these new learning goals, there is an urgent need for knowledge-in-use assessments (see Pellegrino et al., 2014).

A New Vision for K-12 Science Education

The Framework for K-12 Science Education (*Framework*) presents a new vision for science education in which students are to make sense of phenomena or design solutions to problems using disciplinary core ideas, scientific and engineering practices, and crosscutting concepts. This perspective has been referred to as three-dimensional learning (Pellegrino et al., 2014). Disciplinary core ideas (DCIs) represent the powerful ideas of the disciplines of Earth and space sciences, physical science, and life science, and are used in explaining a range of natural phenomena. For instance, the physical science DCI of *matter and its interactions* helps to explain what everything is made of and predicts why things happen in the natural world. Within biology, *evolution* serves as a DCI that explains the diversity of life on Earth. Crosscutting concepts, such as *patterns*, *cause and effect*, *scale*, and *systems* are ideas that occur within and across disciplinary boundaries and have explanatory value throughout much of science and engineering. Patterns, for instance, exist everywhere and occur in biological, chemical, and Earth systems and scientists in all fields seek explanations for observed patterns as they make sense of phenomena. Scientific and engineering practices are the everyday ways of knowing and doing which scientists and engineers employ to study and explore the natural and designed worlds. Both scientists and engineers engage, for example, in the practice of *developing and using models*. Scientists use models to understand and explain phenomena; engineers use models to develop and analyze systems as well as develop and test designs. The *Framework* vision, derived from a rich research base on how students learn science (see, e.g., NRC, 2007), puts forth that in order to learn science, you need to do science by making use of all three dimensions. It is making use of the three dimensions that reflects the knowledge-in-use perspective within the *Framework* and that guided development of the NGSS.

Accordingly, the NGSS expresses standards as *performance expectations* that integrate all three dimensions of science proficiency. Each NGSS performance expectation integrates a science or engineering practice, a disciplinary core idea, and a crosscutting concept into a single statement of *what is to be assessed* at the end of a grade level or grade band. It incorporates all three dimensions of knowledge-in-use by asking students to apply disciplinary knowledge and make connections to a crosscutting concept as they engage in a science or engineering practice to make sense of phenomena or design solutions to problems. For example, an NGSS performance expectation for middle school physical science that focuses on the important idea of chemical reactions is stated as follows: *Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred.* Another performance expectation related to chemical reactions addresses different dimensions and is stated as: *Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved*. In NGSS nomenclature, these are referred to as MS-PS1-2 and MS-PS1-5, respectively (See Tables A1 and A2 in the appendix for complete descriptions of these two performance expectations).

Performance expectations are complex and considered summative goals, and therefore need to be learned over time and through a sequence of carefully designed lessons and units. At the elementary level, students are expected to develop proficiency across the year; whereas at the middle and high school levels, proficiency is attained across the grade band. Given the multidimensionality of the performance expectations and their broad scope, it is no easy task for teachers to gauge student progress toward achieving them.

Designing Classroom-Based Assessment Tasks That Meet the New Vision

Performance expectations present new challenges for anyone involved in science assessment design (Pellegrino et al., 2014): *How do we measure the integration of the three dimensions*? *How can we design integrated assessment tasks in which students make sense of phenomena or design solutions to problems so that they provide evidence of three-dimensional learning?* Additional challenges arise for those concerned with classroom-based assessment design: *How do we use performance expectations in order to construct assessment tasks that can be used during instruction? How can we design these tasks so that they help teachers gauge students' progress toward achieving the performance expectations?* Our approach to addressing both sets of challenges is to use principles of ECD (Almond, Steinberg, & Mislevy, 2002; Mislevy & Haertel, 2006) that have been used in wide-ranging assessment design contexts (e.g., National Center and State Collaborative, 2013; Partnership for Assessment of Readiness for College and Careers, 2014; Smarter Balanced Assessment Consortium, 2012). However, to date, little work has been done on using ECD to design knowledge-in-use assessments for science classroom settings. The need for a principled approach to assessment design, such as ECD, and a focus on classroom assessment, were explicitly discussed in the NRC's report on developing assessments aligned to the NGSS (Pellegrino et al., 2014). We add to the knowledge base by using ECD to apply an evidentiary approach to design assessment tasks that focus on students making sense of phenomena or designing solutions to problems using the three NGSS dimensions.

Our Evidence-Centered Design Approach

To address the goal of developing classroom-based assessment tasks that focus on knowledge-in-use, we use ECD to

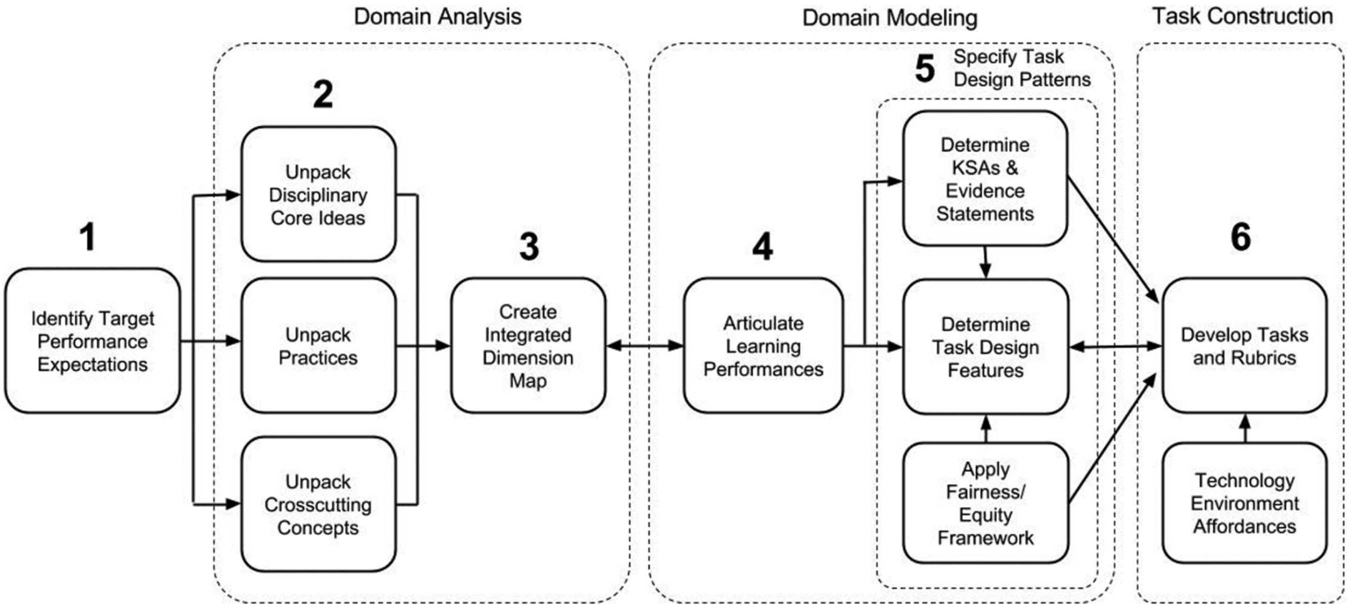

Domain Analysis
Domain Modeling
Task Construction
1
Identify Target Performance Expectations
2
Unpack Disciplinary Core Ideas
Unpack Practices
Unpack Crosscutting Concepts
3
Create Integrated Dimension Map
4
Articulate Learning Performances
5 Specify Task Design Patterns
Determine KSAs & Evidence Statements
Determine Task Design Features
Apply Fairness/ Equity Framework
6
Develop Tasks and Rubrics
Technology Environment Affordances

FIGURE 1. Design process for constructing knowledge-in-use assessment tasks.

systematically unpack NGSS performance expectations and synthesize the unpacking into multiple components that we call learning performances. The term learning performance draws from the work of Perkins (1998) and his notion of understanding performances as opportunities for students to showcase understanding through thought-demanding ways. It has been used more recently in curriculum and assessment design (e.g., DeBarger, Penuel, Harris, & Kennedy, 2015; Krajcik, McNeill, & Reiser, 2008). In our current work, learning performances constitute knowledge-in-use statements that incorporate aspects of disciplinary core ideas, science practices, and crosscutting concepts that students need to be able to integrate as they progress toward achieving performance expectations. A single learning performance is crafted as a knowledge-in-use statement that is smaller in scope and partially represents a performance expectation. Each learning performance describes an essential part of a performance expectation that students would need to achieve at some point during instruction to ensure that they are progressing toward achieving the more comprehensive performance expectation. They collectively describe the proficiencies that students need to demonstrate in order to meet a performance expectation.

Our design process, summarized in Figure 1, enables us to develop classroom-based science assessment tasks that integrate the three dimensions. This process allows us to derive a set of learning performances from a performance expectation or clustered set of performance expectations in a principled way. We use the learning performances to guide the development of assessment tasks and evidence statements. Our design process involves six steps across three distinct phases: (1) *domain analysis*, which involves unpacking of the three NGSS dimensions in the performance expectations to understand the assessable components, (2) *domain modeling*, which involves constructing learning performances and specifying design patterns for tasks associated with them, and (3) *task construction*, using design patterns to create tasks and accompanying rubrics. In this next section, we describe our design process in greater detail and use the performance expectation cluster of MS-PS1-2 and MS-PS1-5 to demonstrate our steps. Of note is that the process we describe has been systematically applied to multiple physical and life science performance expectations for middle school. The resultant tasks (over 100) and related instructional support materials are publicly available through an online portal (https://ngss-assessment.portal.concord.org). After describing and illustrating the design process, we briefly discuss some of the forms of evidence we have collected to validate the general structure of the process and the products produced at different stages in its execution.

Domain Analysis: Unpacking the Dimensions of Performance

In ECD, domain analysis typically entails gathering substantive information about how knowledge is acquired and used in a given domain, such as physical science or life science. After target performance expectations are identified, the process begins with a purposive domain analysis of the three NGSS dimensions that comprise the performance expectations. Because the dimensions are substantially different from the structure of prior standards, a careful and thorough domain analysis is essential to ensure that assessment designers have a deep understanding of them. The resources that we use for unpacking include the *Framework*, NGSS and NGSS appendices, and research literature on the dimensions and their components. Unpacking the dimensions of the target performance expectation(s) is the foundational step in our design approach, as it provides the anchors constituting each dimension and provides a clear focus for what should be elicited in assessment tasks. We use the elaborations from the unpacking to create comprehensive integrated dimension maps that provide a visual representation of the target performance expectations. In Table 1, we provide excerpts from our unpacking of three NGSS dimensions related

Table 1. Excerpts From the Unpacking of Three NGSS Dimensions

Unpacking Aspects of a *Disciplinary Core Idea* Related to Chemical Reactions	
Aspect of a disciplinary core idea	• In a chemical process, the atoms that make up the original substances are regrouped into different molecules, and these new substances have different properties from those of the reactants.
Elaborating the Meaning of Key Sub-Ideas	• *Original substances* change during a chemical process and form new substances with different atomic groupings. • *Atoms are regrouped* and therefore the total number of atoms does not change.
Defining expectations for understanding	• Students should learn (1) that different substances have different atomic groupings, (2) chemical reactions produce new substances with atomic groupings that are different from the original substances, and that (3) mass is conserved in a chemical reaction because the total number of atoms does not change.
Identifying assessment boundaries	• Students are not expected to balance symbolic equations, use atomic masses, or discuss intermolecular forces among atoms.
Prerequisite knowledge	• Knowledge that all matter is made up of particles that are too small to see directly. • Particular materials can be identified by their properties.
Student Challenges	• Students often believe that the total mass decreases during a chemical reaction when a gas is produced (e.g., Nussbaum, 1985).
Relevant Phenomena	• Everyday examples of reactions include combustion (e.g., burning of wood, sugar, steel wool), decomposition reactions (e.g., rotting of bananas and electrolysis of water into oxygen), and mixing (e.g., acid–base reactions). • Pure substances are made from a single type of atom or molecule; examples include sugar (sucrose), sodium chloride, carbon dioxide, oxygen, ammonia, and water.
Unpacking the *Science Practice* of Developing and Using Models	
Aspects of the Practice	• Model elements: Specify or identify elements of the model (and their attributes) and describe why these elements are necessary. • Relationships: Represent or describe the relationships or interactions among model elements and describe why these relationships are important. • Correspondence: Represent or describe the correspondence between model elements/attributes and the target phenomenon or available data. • Limitations: Specify or identify the limitations of the model and describe why these limitations exist.
Intersections with Other Practices	• Models can be used as evidence for explanations and arguments. • Scientific arguments critique or defend the quality or appropriateness of models. • Models can be developed based on results of data analysis. • Investigations may inform the development of models or involve the use of models.
Evidence Required to	• Specifies only the appropriate and necessary elements (and their attributes) in the model needed to explain the target phenomenon and describes why these elements are necessary.
Demonstrate Practice	• Represents only the appropriate relationships and/or interactions among the elements in the model needed to explain the target phenomenon and describes why these relationships are important. • Represents the correspondence between model elements and the real world phenomenon or available data.

(*Continued*)

Table 1. Continued

	• Specifies the appropriate limitations of the model with respect to explaining the target phenomenon and describes why these limitations exist.
Prerequisite Knowledge	• Knowledge that a model contains elements (observable and unobservable) that represent specific aspects of real world phenomena. • Knowledge that models are used to help explain or predict phenomena.
Student Challenges	• Students often view models as physical replicas of objects rather than having explanatory or predictive power (Grosslight, Unger, Jay, & Smith, 1991). • Students often struggle to develop and use models to represent phenomena that are too small to observe directly (Wu, Krajcik, & Soloway, 2001).
Unpacking the *Crosscutting Concept* of Patterns	
Key Aspects	• Ability to identify the presence of patterns in phenomena or data. • Ability to characterize the strength, direction, or nature of patterns in phenomena or data. • Ability to classify objects or relationships into types according to similarities or differences. • Ability to describe why patterns exist and exhibit specific characteristics.
Intersections with Practices	• Explanations address how and why particular patterns occur. • Models describe observed patterns or predict patterns. • Data analysis serves to identify and characterize patterns.
Evidence Required to Demonstrate Application	Students must demonstrate that they can identify, characterize, classify, and describe the reason for the occurrence of three types of patterns: • Repeated occurrences, such as spatially or temporally repeating objects or entities (e.g., extended atomic structures; phase changes). • Similarities, differences, and comparisons of (1) amount or degree across quantities or properties and (2) categories/types of entities: (e.g., comparing physical properties before and after substances interact; distinguishing states and types of matter). • Correlations and trends, such as positive and negative, linear and nonlinear, strong, and weak (e.g., relating particle motion, temperature, kinetic energy, changes in thermal energy, and amount of substance).
Prerequisite knowledge	• Knowledge that patterns are regularly occurring shapes or structures and repeated events, or relationships that can be used to classify objects or attributes. • Knowledge about the characteristics of specific types of patterns, such as the frequency of a repeating event or the strength of a correlation between two variables. • Relevant disciplinary knowledge needed to identify, characterize, and explain observed patterns.

to aspects of the MS-PS1-2 and MS-PS1-5 performance expectations.

Unpacking the Disciplinary Core Ideas. The process of unpacking the disciplinary core ideas that are associated with an NGSS performance expectation or a cluster of performance expectations entails thoughtful consideration of ideas in relation to students' grade level or expected level of expertise. We elaborate the meaning of key sub-ideas, define clear expectations for what ideas students would be expected to use, demarcate boundaries for what students are or are not expected to know, identify background knowledge that is expected of students in order to develop a grade-level-appropriate understanding of aspects a disciplinary core idea, and identify research-based problematic student ideas. We also identify phenomena that provide compelling examples of the disciplinary core idea. Table 1 includes excerpts from our unpacking of aspects of the disciplinary core idea of *matter and its interactions* that relate to chemical reactions as encompassed in MS-PS1-2 and MS-PS1-5.

Unpacking the Science and Engineering Practices. This involves clearly articulating the essential grade-band-appropriate performance for each practice. We articulate specific aspects of practices students are to perform, specify the evidence required for students to demonstrate a high level of proficiency with a practice, identify prior knowledge that is required of students to demonstrate the practice, and identify common challenges that students may encounter as they are developing sophistication with the practice. We also identify productive intersections between the practice and other science practices. Table 1 provides a brief example of unpacking the science practice of *developing and using models*, which is the practice for MS-PS1-5.

Unpacking the Crosscutting Concepts. We identify the important aspects of each, as well as how the crosscutting concepts intersect with targeted science practices and within a particular set of disciplinary core ideas. Similar to our unpacking of practices, we also specify the evidence required for a student to demonstrate a high level of proficiency with the crosscutting concept. Table 1 illustrates unpacking the crosscutting concept of *patterns*, which is the crosscutting concept for MS-PS1-2.

Creating Integrated Dimension Maps. We use the elaborations from the unpacking process to develop integrated dimension maps that lay out the dimensional "terrain" for fully achieving each performance expectation. The maps are visual representations that describe the essential disciplinary core idea relationships and link them to aspects of the targeted crosscutting concepts and science practices (or to closely related crosscutting concepts and practices as identified by the unpacking process). Each map illustrates how the three dimensions work together to define proficiency with a performance expectation and, importantly, shows a range of possible ways to combine aspects of the three dimensions in an assessment. These maps are essential to the principled articulation of three-dimensional learning performances that coherently represent the target performance expectations.

To create an integrated dimension map, we first use the unpacking of one or more disciplinary core ideas to develop a representation that illustrates the relationships between aspects of the core ideas, much like a concept map. The relationships between aspects are then linked via appropriate crosscutting concepts and science practices. In this way, the layout of the aspects associated with the DCI provides an organizing structure for considering how crosscutting concepts and science practices will work together with aspects of the core ideas to represent the breadth of the performance expectation. Once all three dimensions are brought together in a visual representation (i.e., integrated dimension map), we use the map to help specify the learning performances. Figure 2 shows an example of a high-level integrated dimension map for MS-PS1-2 and MS PS1-5.

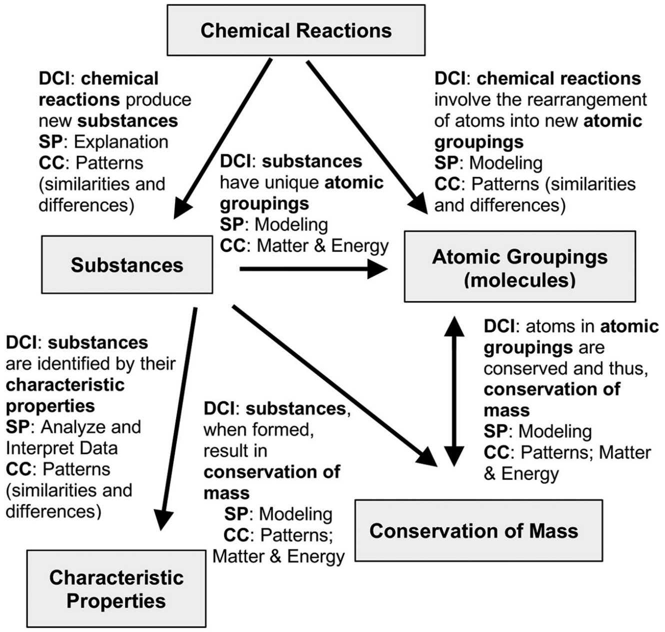

Chemical Reactions
DCI: chemical reactions produce new substances
SP: Explanation
CC: Patterns (similarities and differences)
DCI: chemical reactions involve the rearrangement of atoms into new atomic groupings
SP: Modeling
CC: Patterns (similarities and differences)
DCI: substances have unique atomic groupings
SP: Modeling
CC: Matter & Energy
Substances
Atomic Groupings (molecules)
DCI: substances are identified by their characteristic properties
SP: Analyze and Interpret Data
CC: Patterns (similarities and differences)
DCI: substances, when formed, result in conservation of mass
SP: Modeling
CC: Patterns; Matter & Energy
DCI: atoms in atomic groupings are conserved and thus, conservation of mass
SP: Modeling
CC: Patterns; Matter & Energy
Conservation of Mass
Characteristic Properties

FIGURE 2. Integrated dimension map for MS-PS1-2 and MS-PS1-5.

Domain Modeling: Specifying Knowledge-In-Use Design Patterns

In domain modeling, we consider relationships among the claims we want to make about what students know and can do, evidence that would demonstrate competency with respect to these claims, and features of tasks to elicit the desired evidence. Our claims, evidence, and task features reflect a knowledge-in-use perspective in that we emphasize the application of core ideas and crosscutting concepts through engagement in a science practice. The learning performances we specify serve as the claims, as they clearly specify what is expected of students to demonstrate in order to provide evidence that they have achieved one or more aspects of a performance expectation. Learning performances represent a keystone in the evidence-based argument that our assessment tasks represent the NGSS performance expectations for formative assessment purposes.

Articulating Learning Performances. We use the integrated dimension map in tandem with our unpacking to articulate and refine a set of learning performances that collectively describe the proficiencies that students need to demonstrate in order to meet a performance expectation (see Table 2). Together, a set of learning performances provides the detail needed to create a coherent set of assessment tasks that can provide evidence that students can use and apply the knowledge aligned to a performance expectation or cluster of performance expectations. Learning performances are akin to learning goals that take on the three-dimensional structure of the performance expectations—they articulate and integrate assessable aspects of performance that build toward the more comprehensive NGSS performance expectation.

Creating Design Patterns. For each learning performance, we specify a design pattern (Mislevy & Haertel, 2006) that

Table 2. Two NGSS Performance Expectations and Related Learning Performances

MS-PS1-2: Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred.

MS-PS1-5: Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved.

Learning Performances (LPs) for MS-PS1-2 and MS-PS1-5

LP 1: Analyze and interpret data to determine whether substances are the same based upon patterns in characteristic properties.

LP 2: Construct a scientific explanation about whether a chemical reaction has occurred using patterns in data on properties of substances before and after the substances interact.

LP 3: Use reasoning from patterns to evaluate whether a model explains that a chemical reaction produces new substances and conserves atoms.

LP 4: Use a model to explain that in a chemical reaction, atoms are regrouped and this is why mass is conserved.

LP 5: Develop a model of a chemical reaction that explains that new substances are formed by the regrouping of atoms and that mass is conserved.

LP 6: Use reasoning about matter and energy to evaluate whether a model explains that a chemical reaction produces new substances and conserves mass because atoms are conserved.

guides the principled development of tasks. Design patterns include numerous elements to ensure that the tasks elicit evidence of proficiency with the learning performance: *Focal Knowledge, Skills, and Abilities* (KSAs) refer to the proficiencies to be targeted by the assessment task. We articulate multiple KSAs for a learning performance to capture the range of proficiencies needed to demonstrate that learning performance. *Evidence statements* articulate the observable features of student performance that can provide evidence of a high-level demonstration of the learning performance and we use these to inform the development of both tasks and scoring rubrics. *Characteristic task features* describe the attributes that are common across all the tasks for a learning performance. *Variable task features* describe the features that can vary across tasks, such as the level of scaffolding to vary task difficulty. Both types of task features include *equity/fairness considerations* to help ensure that our tasks are accessible and fair to students of diverse cultural, linguistic, and socioeconomic backgrounds. To articulate these task design features, we use an equity/fairness framework that draws from Universal Design for Learning (UDL) (Rose & Meyer, 2006; Rose, Meyer, & Hitchcock, 2005)—which articulates a set of guiding principles for designers to accommodate individual differences—and is informed by research on fair and equitable assessment practices in science (e.g., Lee, Quinn, and Valdés, 2013; Wolf & Leon, 2009). Table 3 illustrates a design pattern for Learning Performance 5 as articulated in Table 2 above.

Task Construction: Developing Three-Dimensional Tasks and Rubrics

The final phase of the design process involves using the design patterns to construct assessment tasks and rubrics aligned with each learning performance. The task designs make use of both characteristic and variable task features, allowing for the development of multiple tasks within a 'family' that vary in difficulty level while maintaining alignment with the learning performance. The task design process also considers the ways student responses will be scored and evaluated for evidence of the focal KSAs. As assessment tasks are developed, we use the focal KSAs and evidence statements from the design patterns to develop a scoring rubric for each.

We intend our tasks to be used flexibly by teachers during instruction and, accordingly, the tasks we design are relatively short in duration, requiring anywhere from 5 to 10 minutes to complete depending on the requirements of particular tasks. Each task is anchored in a phenomenon and most are contextualized within a brief scenario. The tasks are technology-based and made available online through a web-based portal. Teachers select the tasks they would like to use from the online task portal and decide how they would like to use the tasks with their students. During the development of tasks, we build in specific technology enhancements, such as simulations and drawing tools, to support students in engaging with all three dimensions of the learning performance.

Task Example. Figure 3 illustrates a physical science assessment task designed for classroom use. It was developed to assess a learning performance aligned with the NGSS performance expectation MS-PS1-5. The task is one drawn from a set of tasks developed using the design pattern for learning performance 5 (see Table 2), which emphasizes reasoning about interactions of matter to develop a model that shows that a chemical reaction regroups and conserves atoms. The prompts within the task (i.e., draw and use a model) align to the evidence statements in the design pattern and, in this way, we ensure that the task elicits evidence of a student's three-dimensional proficiency with the learning performance.

Rubric Example. The evidence statements also constitute the basis for a scoring rubric for each assessment task, ensuring coherence among the task design, scoring approach, and desired evidence of students' proficiency. The rubric development approach centers on the development of multiple rubric components. Each rubric component measures proficiency with a specific focal KSA and builds from its corresponding evidence statement. Table 4 illustrates an example of a scoring rubric for the Battery in Water task (Figure 3). The rubric components address the evidence statements associated with the task. When scoring a student's response to a task, each rubric component is applied to the response, obtaining a set of scores that collectively describe the student's proficiency with the learning performance. In this rubric, which was designed for purposes of task validation, the components are scored polytomously from 0 to 2. In addition to separating multiple aspects of a student's proficiency needed to respond

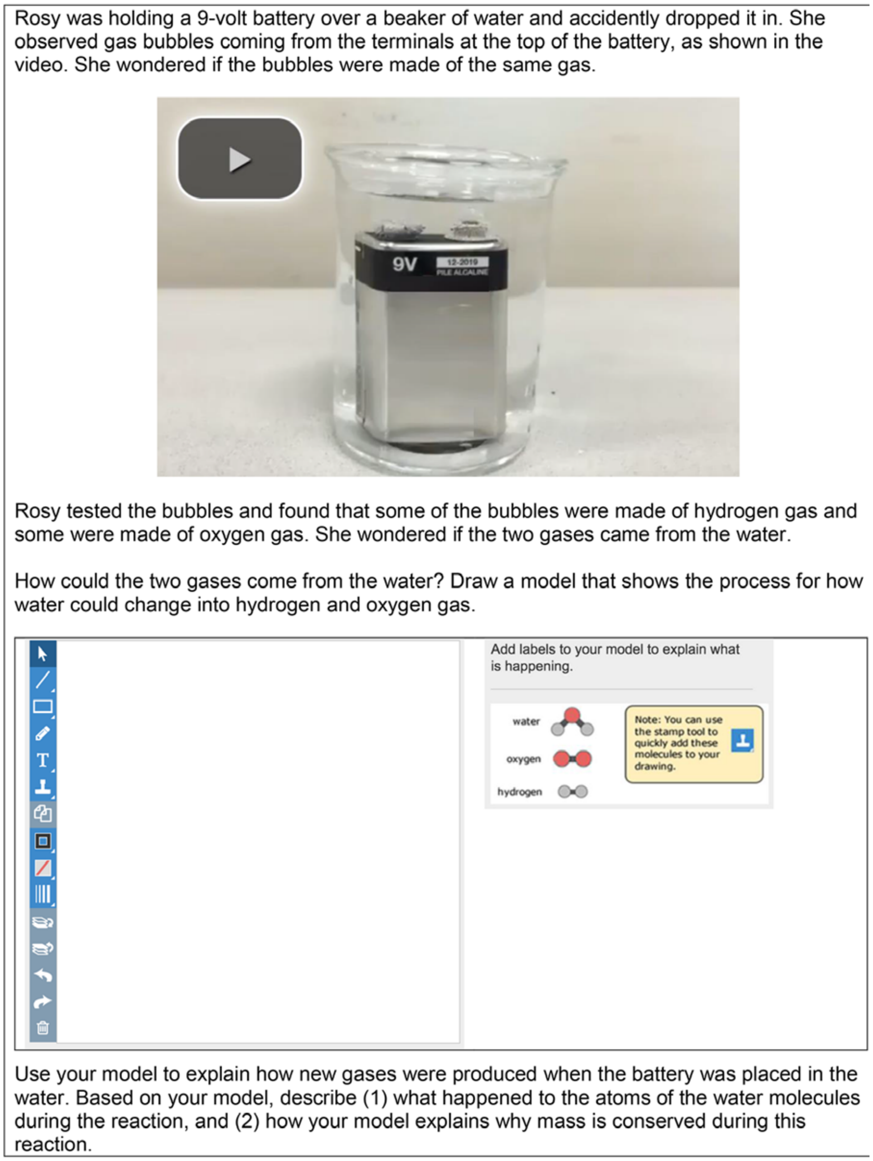
Table 3. Knowledge-in-Use Design Pattern for a Learning Performance

Learning Performance (Claim)	• Learning Performance 5: Develop a model of a chemical reaction that explains that new substances are formed by the regrouping of atoms and that mass is conserved.
Focal Knowledge, Skills, and Abilities (KSAs)	• Apply the scientific principle that chemical reactions produce new substances with atomic arrangements that are different from the original substances. • Apply the scientific principle that mass is conserved in chemical reactions because the total number of atoms does not change. • Develop a model with correspondence between model features and atoms rearranging in a chemical reaction to produce new substances and conserve atoms. • Support model use by explaining that in a chemical reaction, atoms rearrange (regroup) and are conserved.
Evidence Required to Demonstrate Proficiency	• A model of a chemical reaction that shows that atoms are correctly regrouped from reactants to products and that conserves each type of atom. • A statement of how many of each type of atom are shown in the model before and after the process occurs. • A statement describing that the atoms in the model are regrouped during the process and thus conserved in a chemical reaction.
Characteristic Task Features	• The term "substance" means a pure substance (not a mixture of substances). • A description of a scenario involving a simple chemical reaction that is accessible to a broad range of students. • A drawing tool enabling students to represent the reactants and products of a chemical reaction at the atomic level. • A prompt to draw a model of the chemical reaction. • A prompt to use a model to explain what happens to the atoms during the reaction. • Use of straightforward language that is accessible to students with diverse linguistic abilities.
Variable Task Features	• Types of models used to illustrate chemical reactions. • State of the substances in question (i.e., solid, liquid, or gas state). • Types, numbers, and arrangements of atoms and/or molecules included as data in the model. • Task scaffolding features to help elicit relevant data patterns and scientific principles. • Visual aids to support comprehension by students with diverse linguistic and visual processing abilities.

correctly to a task, the individual rubric components focus scorers' attention on specific features of a student's response, thereby promoting reliability in scoring. While this rubric was created for research purposes, we also design rubrics in a form that is both practical for classroom use and educative for teachers about the nature of the NGSS. From the rubric illustrated in Table 4, for instance, we can derive a classroom rubric that teachers can use for formative assessment. These classroom rubrics can be designed to facilitate rapid scoring judgments to provide teachers with timely insights about their students' progress. They can also phrase scoring criteria in ways that highlight for teachers how to evaluate students' proficiency with integrating the three NGSS dimensions. We have done this strand of work with input from teachers to facilitate their interpretation of key aspects of student performance.

Validation of the Design Process and Products

In developing, refining, and applying our design process to multiple performance expectations in the physical and life sciences, we have been very cognizant of generalization and



Rosy was holding a 9-volt battery over a beaker of water and accidently dropped it in. She observed gas bubbles coming from the terminals at the top of the battery, as shown in the video. She wondered if the bubbles were made of the same gas.

Rosy tested the bubbles and found that some of the bubbles were made of hydrogen gas and some were made of oxygen gas. She wondered if the two gases came from the water.

How could the two gases come from the water? Draw a model that shows the process for how water could change into hydrogen and oxygen gas.

Add labels to your model to explain what is happening.

Note: You can use the stamp tool to quickly add these molecules to your drawing.

Use your model to explain how new gases were produced when the battery was placed in the water. Based on your model, describe (1) what happened to the atoms of the water molecules during the reaction, and (2) how your model explains why mass is conserved during this reaction.

FIGURE 3. Example physical science assessment task: Battery in Water. [Color figure can be viewed at wileyonlinelibrary.com]

validation issues. Our validation efforts have been guided by the framework presented by Pellegrino, DiBello, and Goldman (2016) for instructionally supportive assessments. That framework discusses three major components of validity for which systematic evidence should be sought above and beyond articulation of the ECD process: Cognitive, Instructional, and Inferential. While it is beyond the scope of this paper to discuss the multiple forms of evidence that have been obtained to validate the process and the products of key stages of the process, we can briefly summarize some of that here.

For example, at each major stage in the process, we conduct an independent review of the products of the domain analysis process by having science and science education experts review the integrated dimension maps and the learning performances derived from them. This includes the appropriateness of each designated learning performance and the adequacy of the set of learning performances with respect to representation and coverage of the domain. Based on feedback from the experts, we make revisions or clarifications as needed. We also ask these same experts to review the tasks that we designed to align with each learning performance and review the accompanying evidence rules and scoring rubrics, again making refinements in response to feedback. Throughout the process, we conduct an equity/fairness review to ensure that tasks minimize bias. Once we have tasks that have been through the expert review phases, we further refine them using several steps, including (1) think-aloud sessions with students that examine whether tasks are comprehensible to them and whether they elicit three dimensional performance, (2) collection of classroom performance data to determine applicability and reliability of scoring rules

Table 4. Example Rubric for the Battery in Water Task

Score	Rubric Component (1)	Rubric Component (2)	Rubric Component (3)
	Draws a model whose atoms are correctly regrouped from reactants to products and that conserves each type of atom	States how many of each type of atom are shown in the model before and after the process occurs	States that the atoms in the model are regrouped during the process
2	Student model shows • Equal number of atom types on each side (e.g., 2 O atoms and 4 H atoms on each side) AND • H2O as the reactant and H2 and O2 as the products	Student states • Both sides have equal number of atom types (e.g., 2 O atoms and 4 H atoms) AND/OR • Mass is conserved because the total number of atoms does not change	Student states that reaction products are formed in the following way: • Reactant molecules break apart and the atoms are regrouped to form product molecules
1	Student model shows • Equal numbers of O and H atoms on each side OR • H2O as the reactant and H2 and O2 as the products	Student states • Both sides have equal number of O or H atoms (e.g., 2 O atoms OR 4 H atoms on each side) AND/OR • Atoms are neither created nor destroyed	Student states that reaction products are formed in at least one of the following ways: • Atoms are regrouped or rearranged • Reactant molecules break apart and form product molecules
0	Student model includes • Unequal numbers of O and H atoms on each side AND/OR • Reactants/products other than H2O, H2, or O2	Student response includes • Missing/incorrect statement about atom conservation	Student response includes: • Missing/incorrect statement about regrouping OR • Refers to breaking apart or formation but not both

using the rubrics, (3) application of measurement models to the scored data to examine item performance characteristics, and (4) classroom studies with teachers, who provide design feedback on tasks and help us consider the possibilities for formative use. More detailed discussions of specific validation activities and results can be found in several papers (e.g., Alozie et al., 2018; Gane, McElhaney, Zaidi & Pellegrino, 2018; McElhaney, Zaidi, Gane, Alozie, & Harris 2018; Zaidi et al., 2018).

The Nature of Our Design Work

We acknowledge that there are many different ways to develop and embed formative assessment in instruction, such as developing comprehensive performance tasks that require extended periods of time and involving instructional activities. Our work, on the other hand, entails developing tasks at a much smaller grain size that is usable in the course of instruction to provide information about students' development toward achieving a given performance expectation (or cluster). The shorter task length balances the desire to engage students in authentic science practices with the need for teachers to use the tasks flexibly during instruction and get timely information from the tasks for formative purposes. In our approach, a task or scenario might include various types of student responses including written elements, drawings, or interactions with simulations. We intend for teachers to use several tasks at appropriate points during instruction to gauge their students' progress toward achieving the performance expectations.

Considerations in Using Our Design Approach

Our design approach has several important advantages given the nature of knowledge-in-use learning goals and the need for assessments that can be instructionally supportive. First, it reflects a broadly accessible vision of how to design knowledge-in-use assessments and provides a systematic approach for documenting principled design decisions. We used science in this article as an illustrative case. Our use of this approach for science is not discipline or grade-band specific—we expect it to generalize from our initial work in middle school physical and life sciences to the other science and engineering disciplines and grade-bands—and we are presently using the approach to develop tasks and accompanying rubrics that address performance expectations in the elementary grades. We also expect that our approach can be applied in other domains such as mathematics or history where there are expectations that involve reasoning with and about content using disciplinary practices. Second, our ECD-based process, because of its explicit focus on domain analysis and domain modeling, adds consistency to the design work so that designers can develop a variety of tasks that share common design elements and all align to the target knowledge-in-use learning goals. Ensuring consistency and alignment is a major challenge shared throughout the assessment community, particularly in the era of designing assessments of knowledge-in-use. Third, as illustrated in Figure 2, the creation of integrated dimension maps allow for a systematic process of linking the domain analysis with the domain modeling through the articulation of learning performances. The notion of unpacking is not new, but how a designer can adequately represent the complexity and link the unpacking to claims and design patterns (the design modeling) has never been clearly articulated in previous literature. Fourth, the articulation of a set of learning performances allow for the design of classroom-based tasks that align to more comprehensive knowledge-in-use learning goals. These tasks represent aspects of large grain-size learning goals that students need to meet as they progress in their learning. Thus, tasks constructed from learning performances can provide actionable information on how students are making progress toward the intended learning goals.

While our design approach has important advantages, challenges also exist. The unpacking of the dimensions requires careful attention and sufficient expertise among team members to firmly understand the three dimensions and how they work individually and together. Incomplete or errant unpacking can disrupt the entire design process. A core practical challenge is to ensure that developed tasks are valid and fair and accessible to a wide range of learners, especially those with different linguistic backgrounds. Using clear and accessible language is paramount, as is the need to select appropriate meaningful phenomena that will be familiar to as wide a range of students as possible. Another challenge is the development of rubrics that provide insight into student learning and are interpretable and useable by classroom teachers. A central question we have wrestled with is whether rubrics should integrate the dimensions into a single score or separately evaluate aspects of performance for each of the three dimensions. We chose to take an integrated approach to interpreting student responses thereby staying true to the *Framework* vision that science proficiency means being able to integrate the three dimensions. From a learning perspective, assessments that integrate key aspects of all three dimensions seem to be feasible and should provide insights into student achievement and its change over time with instruction. Yet we recognize too that other researchers might approach this quite differently. For example, there may be a research need and perhaps very good instructional argument for examining the extent to which each of the three performance dimensions are separable and identifiable.

Summary and Conclusions

The kind of deeper learning emphasized by knowledge-in-use learning goals requires improved assessments with new types of tasks and situations that call upon students to demonstrate well-integrated learning. Accordingly, new design approaches are needed for creating assessment tasks that align with ambitious knowledge-in-use learning goals and that provide teachers with actionable information about how students are progressing toward meeting them. Our design approach offers a useful structure that may help to build coherence in assessments and assessment systems. For example, the investment in unpacking and generating of integrated dimension maps, learning performances, and design patterns provides opportunities for collaboration and shared agreement among designers that can help in establishing coherence and strengthen the overall assessment argument. The approach could potentially be used for any multidimensional performance construct.

While our work is early and ongoing, we see promise for helping to provide answers to critical questions related to the design and use of assessments of knowledge in use. Importantly, our work provides an early-stage example of one way to approach the development of high-quality assessments that elicit knowledge-in-use performance. Having the right kinds of assessments is critically important because they guide what teachers and students attend to during instruction. High-quality assessments can help teachers implement new standards, help students learn more, and provide equitable opportunities for all students to develop their proficiencies within and across disciplines.

Acknowledgments

This work was supported through funding by the National Science Foundation (grant numbers 1316874, 1316903, 1316908, & 1903103) and the Gordon and Betty Moore Foundation (grant number 4482). Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily represent the views of their institutions, the National Science Foundation, or the Gordon and Betty Moore Foundation. The authors gratefully acknowledge the following colleagues for their contributions to the research and development effort: Nonye Alozie, Chanyah Dahsah, Daniel Damelin, Louis DiBello, Reina Fujii, Brian Gane, Diksha Gaur, Mon Lin Ko, Jane Lee, Krystal Madden, Kevin McElhaney, Kiley McElroy-Brown, Consuelo Morales, Phyllis Pennock, Samuel Severance, Gauri Vaishampayan, and Sania Zaidi.

Appendix

Table A1. NGSS Performance Expectation MS-PS1-2

Students who demonstrate understanding can:

MS-PS1-2 Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred. [Clarification Statement: Examples of reactions could include burning sugar or steel wool, fat reacting with sodium hydroxide, and mixing zinc with hydrogen chloride.] [*Assessment Boundary: Assessment is limited to analysis of the following properties: density, melting point, boiling point, solubility, flammability, and odor.*]

The performance expectation above was developed using the following elements from the NRC document A Framework for K-12 Science Education:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Analyzing and Interpreting Data Analyzing data in 6–8 builds on K-5 and progresses to extending quantitative analysis to investigations, distinguishing between correlation and causation, and basic statistical techniques of data and error analysis. • Analyze and interpret data to determine similarities and differences in findings. ***Connections to Nature of Science*** **Scientific Knowledge is Based on Empirical Evidence** • Science knowledge is based upon logical and conceptual connections between evidence and explanations.	**PS1.A: Structure and Properties of Matter** • Each pure substance has characteristic physical and chemical properties (for any bulk quantity under given conditions) that can be used to identify it. **PS1.B: Chemical Reactions** • Substances react chemically in characteristic ways. In a chemical process, the atoms that make up the original substances are regrouped into different molecules, and these new substances have different properties from those of the reactants.	**Patterns** • Macroscopic patterns are related to the nature of microscopic and atomic-level structure.

Connections to other DCIs in this grade-band:
MS.PS3.D ; **MS.LS1.C** ; **MS.ESS2.A**

Articulation of DCIs across grade-bands:
5.PS1.B ; **HS.PS1.B**

Common Core State Standards Connections:
ELA/Literacy -

RST.6-8.1	Cite specific textual evidence to support analysis of science and technical texts, attending to the precise details of explanations or descriptions.(MS-PS1-2)
RST.6-8.7	Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). (MS-PS1-2)

Mathematics -

MP.2	Reason abstractly and quantitatively. (MS-PS1-2)
6.RP.A.3	Use ratio and rate reasoning to solve real-world and mathematical problems. (MS-PS1-2)
6.SP.B.4	Display numerical data in plots on a number line, including dot plots, histograms, and box plots. (MS-PS1-2)
6.SP.B.5	Summarize numerical data sets in relation to their context. (MS-PS1-2)

Note: Adapted from *Next Generation Science Standards: For states, by states* (NGSS Lead States, 2013).

Table A2. NGSS Performance Expectation MS-PS1-5

Students who demonstrate understanding can:

MS-PS1-5 Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved. [Clarification Statement: Emphasis is on law of conservation of matter and on physical models or drawings, including digital forms, which represent atoms.] [*Assessment Boundary: Assessment does not include the use of atomic masses, balancing symbolic equations, or intermolecular forces.*]

The performance expectation above was developed using the following elements from the NRC document A Framework for K-12 Science Education:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Developing and Using Models Modeling in 6–8 builds on K-5 and progresses to developing, using and revising models to describe, test, and predict more abstract phenomena and design systems. • Develop a model to describe unobservable mechanisms. ***Connections to Nature of Science*** **Science Models, Laws, Mechanisms, and Theories Explain Natural Phenomena** • Laws are regularities or mathematical descriptions of natural phenomena.	**PS1.B: Chemical Reactions** • Substances react chemically in characteristic ways. In a chemical process, the atoms that make up the original substances are regrouped into different molecules, and these new substances have different properties from those of the reactants. • The total number of each type of atom is conserved, and thus the mass does not change.	**Energy and Matter** • Matter is conserved because atoms are conserved in physical and chemical processes.

Connections to other DCIs in this Grade-Band:
MS.LS1.C ; **MS.LS2.B** ; **MS.ESS2.A**

Articulation of DCIs Across Grade-Bands:
5.PS1.B ; **HS.PS1.B**

Common Core State Standards Connections:
ELA/Literacy -

RST.6-8.7	Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). *(MS-PS1-5)*
Mathematics -	
MP.2	Reason abstractly and quantitatively. (MS-PS1-5)
MP.4	Model with mathematics. (MS-PS1-5)
6.RP.A.3	Use ratio and rate reasoning to solve real-world and mathematical problems. (MS-PS1-5)

Note: Adapted from *Next Generation Science Standards: For states, by states* (NGSS Lead States, 2013).

References

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*, 1–64.

Alozie, N., Haugabook Pennock, P., Madden, K., Zaidi, S., Harris, C. J., & Krajcik, J. S. (2018). *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.

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

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan*, *80*, 139–148.

Bransford, J. D., Brown, A., Cocking, R., Donovan, M. S, & Pellegrino, J. W. (Eds.). (2000). *How people learn: Brain, mind, experience and school*. (Expanded edition). Washington, DC: The National Academies Press.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, *48*, 1109–1136.

Common Core State Standards Initiative. (2010a). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/ELA-Literacy/

Common Core State Standards Initiative. (2010b). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/Math/

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2015). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, *37*(2), 174–192.

European Commission. (2012). *Rethinking education*. Retrieved from http://ec.europa.eu/education/policy/multilingualism/rethinking-education_en

Finnish National Board of Education (FNBE). (2015). *National Core Curriculum for General Upper Secondary Schools 2015*. Helsinki: Finnish National Board of Education (FNBE).

Gane, B. D., McElhaney, K. W., Zaidi, S. Z., & Pellegrino, J. W. (2018). *Analysis of student and item performance on three-dimensional constructed response assessment tasks*. Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28*, 799-822.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education*, *92*, 1–32.

Kulgemeyer, C & Schecker, H. (2014). Research on educational standards in German science education—Towards a model of student competences. *EURASIA Journal of Mathematics, Science and Technology Education*, *10*, 257–269.

Lee, O., Quinn, H., and Valdés, 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*, 223–233.

McElhaney, K.W., Zaidi, S., Gane, B. D., Alozie, N., & Harris, C.J. (2018, March). *Designing NGSS-aligned assessment tasks and rubrics to support classroom-based formative assessment*. Paper presented at the NARST Annual International Conference, Atlanta, GA.

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

National Center and State Collaborative. (2013). *NCSC alternate assessment.* Minnneapolis, MN: University of Minnesota, National Center and State Collaborative Consortium.

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

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

National Science Teachers Association. (2016). *NSTA position statement: The next generation science standards*. Retrieved from https://www.nsta.org/about/positions/

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Nussbaum, J. (1985). The particulate nature of matter in the gaseous phase. In R. Driver, E. Guesne, & A. Tiberghien (Eds.), *Children's ideas in science* (pp. 124–144). Milton Keynes, UK: Open University Press.

OECD. (2016). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris, France: OECD Publishing.

Partnership for Assessment of Readiness for College and Careers. (2014). *PARCC Model Content Frameworks for Mathematics 2014*. Washington, DC: Author.

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

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*, 59–81.

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

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

Perkins, D. (1998). What is understanding? In M. S. Wiske (Ed.), *Teaching for understanding: Linking research with practice* (pp. 39–58). San Francisco, CA: Jossey-Bass.

Rose, D. H., & Meyer, A. (2006). *A practical reader in universal design for learning*. Cambridge, MA: Harvard Education Press.

Rose, D. H., Meyer, A., & Hitchcock, C. (Eds.). (2005). *The universally designed classroom: Accessible curriculum and digital technologies*. Cambridge, MA: Harvard Education Press.

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*, 57–84.

Smarter Balanced Assessment Consortium. (2012). *Smarter balanced assessments*. Sacramento, CA: Author.

Tucker, C. G. (2015). *Psychometric considerations for the next generation of performance assessment with implications for policy and practice*. Princeton, NJ: Educational Testing Service.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*, 139–159.

Wu, H. K., Krajcik, J., & Soloway, E. (2001). Promoting conceptual understanding of chemical representations: Students' use of a visualization tool in the classroom. *Journal of Research in Science Teaching*, *38*, 821–842.

Zaidi, S. Z., Ko, M., Gane, B. D., Madden, K., Gaur, D., & Pellegrino. J.W. (2018). *Portraits of teachers using three-dimensional assessment tasks to inform instruction*. Paper presented at the NARST Annual International Conference, Atlanta, GA.