Designing Scoring Rubrics to Support NGSS-aligned, Classroom-based Formative Assessment

Paper presented at the AERA Annual Meeting
San Antonio, TX: April 27 - May 1, 2017

Kevin W. McElhaney[1], Brian D. Gane[2], Louis V. DiBello[2], Reina Fujii[1], Phyllis Haugabook Pennock[3], Gauri Vaishampayan[2], James W. Pellegrino[2]

[1]Center for Technology in Learning, SRI International
[2]Learning Sciences Research Institute, University of Illinois at Chicago
[3]CREATE for STEM Institute, Michigan State University

Abstract: We describe how principles of evidence-centered design inform the development of classroom-based science assessment tasks and rubrics that integrate three dimensions of science proficiency addressed in the Next Generation Science Standards (NGSS). Beginning with a specific NGSS performance expectation (PE), our design process involves the articulation of *learning performances*—a set of claims that collectively represent the proficiencies of the target PEs and to which assessment tasks and rubrics are aligned. For each learning performance we specify an assessment design pattern that identifies (a) aspects of proficiency that are of particular importance to classroom-based, formative assessment, (b) student evidence for those proficiencies, and (c) task features that elicit this evidence. Using this design pattern as the basis for both task and rubric design ensures the alignment of tasks and rubrics with the PEs. This paper describes our design approach and includes accompanying examples of assessment design artifacts. We also consider assessment design challenges, next steps, and implications of this work for the next generation of science assessments.

Suggested citation:
McElhaney, K.W., Gane, B. D., DiBello, L.V., Fujii, R., Pennock, P.H, Vaishampayan, G., Pellegrino, J.W. (2017, May). *Designing scoring rubrics to support NGSS-aligned, classroom-based formative assessment*. Paper presented at the American Educational Research Association Annual Meeting, San Antonio, TX.

Introduction

A significant challenge facing science educators who are shifting instruction to align with the *Next Generation Science Standards* (NGSS) (NGSS Lead States, 2013) is how to assess students' progress toward achieving the new standards. The NGSS reflects an ambitious vision for science education presented in the National Research Council's (NRC) *Framework for K-12 Science Education* (NRC, 2012). This vision describes the integration of disciplinary core ideas, crosscutting concepts, and science and engineering practices as fundamental to improving students' understanding and promoting students' participation in science as a career professional or citizen. The *Framework* emphasizes rich science learning as requiring a tight coupling of what students know (content knowledge) and what they can do (practice). In this paper, we describe a principled design process for creating classroom-based science assessments and scoring rubrics aligned with the NGSS performance expectations. The overarching aim of our design work is to create instructionally supportive tasks and rubrics that can be used by teachers to help advance student learning in classrooms implementing the NGSS.

Our design approach responds to challenges identified by Gorin and Mislevy (2013) and Scalise (2014a, 2014b) for next-generation science assessment design, including complex domain definitions, complex student performances, and the increased importance of diagnostic reporting for formative purposes. Using this approach, we have developed assessment tasks and rubrics that (a) integrate the three dimensions of the NGSS at every phase of development, (b) maintain clear links among the target NGSS performance expectations, evidence of student proficiency, task design features, and rubric design, and (c) distinguish multiple aspects of student proficiency in a way that provides insights to teachers about their students' progress and can inform teachers' instructional decisions.

Rationale and Design Framework

Integrating science content knowledge and science practice. The shift to integrating science practices with content knowledge is based on studies of professional scientific practice (e.g., Latour & Woolgar, 1996) and on empirical research conducted in the years since the publication of the *National Science Education Standards* (NRC, 1996) and the *Benchmarks* (American Association for the Advancement of Science, 1993). Selected practices, such as argumentation and modeling, received little attention in these earlier standards documents but are now more prominent (Duschl & Osborne, 2002; Lehrer & Schauble, 2006). Much of this contemporary research is synthesized in reports such as *How People Learn* (NRC, 2000), *Taking Science to School* (NRC, 2007), and *Learning Science in Informal Environments* (NRC, 2009). These reports, as well as the *Framework* (NRC, 2012), contend that proficiency in science requires using and applying knowledge in the context of science practice. When students have opportunities to use science practices to develop, test, and apply their ideas, they deepen their conceptual knowledge as well as their ability to engage in the practices of science. This knowledge-in-use perspective (Harris & Salinas, 2009; Krajcik, McNeill, & Reiser, 2008) as instantiated in the NGSS holds that disciplinary core ideas, science and engineering practices, and crosscutting concepts together enable learners to make sense of phenomena and design solutions to problems. Consequently, in the NGSS these three dimensions are integrated into knowledge-in-use statements called *performance expectations* (PEs).

The complexity of the NGSS PEs presents formidable challenges for assessment design. For example, traditional approaches to assessment design that only target disciplinary content knowledge will be inadequate for creating tasks and rubrics that must measure students' integrated performance (Pellegrino, Wilson, Koenig, & Beatty, 2014). In addition, PEs represent end-of-grade-band performance targets and therefore often incorporate a wide range of proficiencies that may be difficult to assess in a single assessment task. Moreover, designing assessment tasks and rubrics to be used formatively during the course of instruction will require an approach that can decompose PEs in a systematic way while retaining their three-dimensional nature. Addressing these challenges requires a principled design process that allows assessment developers to align classroom-based assessment tasks and rubrics to PEs.

Designing for formative assessment. Our design approach builds from research illustrating the benefits of classroom-based formative assessment (e.g., Black & Wiliam, 1998; Kingston & Nash, 2011). In formative assessment, teachers question students or engage them in activities that provide evidence of their proficiency with learning goals. Teachers may use this evidence to inform their subsequent instructional decisions. Ruiz-Primo and Furtak (2007) describe a range of classroom-based formative assessment types as existing along a spectrum from formal to informal. Formal formative assessment entails using previously planned activities or tasks designed to elicit specific aspects of student proficiency related to the instructional targets, while informal formative assessment approaches are implemented spontaneously, typically by teacher questioning or group discussion. Our assessment tasks and rubrics are aimed to support formal approaches to formative assessment. As such, our tasks and rubrics are aligned with specific aspects of proficiency with an NGSS PE and are designed to be used to gauge student progress toward these PEs during the course of instruction.

Evidence-centered design (ECD) (Mislevy & Haertel, 2006) provides a conceptual framing for analyzing content for assessment design, and it can be used to identify, organize, and document these essential and assessable components of NGSS PEs (Harris, Krajcik, Pellegrino, & McElhaney, 2016). ECD-based design patterns explicate an argument about what inferences about student proficiency can be made based on evidence found in students' work products. Design patterns structure the linkages among the targeted student proficiencies, assessment task design features, observable evidence, and scoring. Design patterns describe the kinds of assessment tasks that elicit target constructs and demonstrate how particular performances provide evidence for students' knowledge, skills, and abilities (Mislevy & Haertel, 2006; Songer, Kelcey, & Gotwals, 2009). The up-front specification of the design framework via ECD promotes systematicity in assessment task and rubric design.

To address the goals of formative assessment during the course of instruction, we use ECD to systematically decompose PEs into multiple *learning performances* that can guide formative assessment opportunities (DeBarger et al., 2014; Harris, McNeill, Lizotte, Marx, &

Krajcik, 2006; Krajcik, McNeill, & Reiser, 2008). Learning performances are knowledge-in-use statements that incorporate aspects of disciplinary core ideas, science and engineering practices, and crosscutting concepts that students need to attain as they progress toward achieving proficiency with a PE. Learning performances are akin to learning goals that take on the three-dimensional structure of the PEs—they articulate and integrate assessable aspects of performance that build toward the more comprehensive PE. For classroom purposes, the learning performances also help identify important formative assessment opportunities for teachers. Our design process enables us to derive a set of learning performances from a PE in a principled way (described below) that ensures the learning performances meet these requirements.

Assessment Task and Rubric Design Approach

The design process (Figure 1) emphasizes three primary ECD phases, each involving multiple components. In the domain analysis phase, we unpack the three dimensions of the NGSS PEs and explicitly represent the essential relationships in the domain. In the domain modeling phase, we derive a set of learning performance statements from the domain analysis. Additionally, for each learning performance, we articulate design patterns that specify the foundation upon which to build tasks and rubrics (Mislevy & Haertel, 2006). In the task and rubric development phase, design patterns and technology affordances of the task delivery system inform the development of both tasks and rubrics in a way that aligns the assessment targets, desired evidence of student proficiency, task design features, and scoring criteria.

This design process can address either a single PE or a coherent bundle of PEs. In this paper, we will use examples from our design documentation for tasks and rubrics addressing the NGSS PE **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.** This PE integrates disciplinary core ideas related to chemical reactions, the science practice of developing and using models, and the crosscutting concept of energy and matter.

Domain Analysis Phase

In ECD, domain analysis entails gathering substantive information about how knowledge and skills are acquired and used in the domain for the purpose of designing assessments. In our design process, the domain analysis guides the eventual articulation of learning performances associated with the target PE. This domain analysis involves (1) unpacking the disciplinary core ideas, science and engineering practices, and crosscutting concepts that are related to the target performance expectations and (2) constructing an integrated dimension map that describes the essential disciplinary relationships and links them to aspects of the targeted crosscutting concepts and science and engineering practices. These steps, briefly described below, are elaborated in detail elsewhere (Harris et al., 2016; McElhaney, Gane, Harris, Pellegrino, DiBello, Krajcik, 2016).

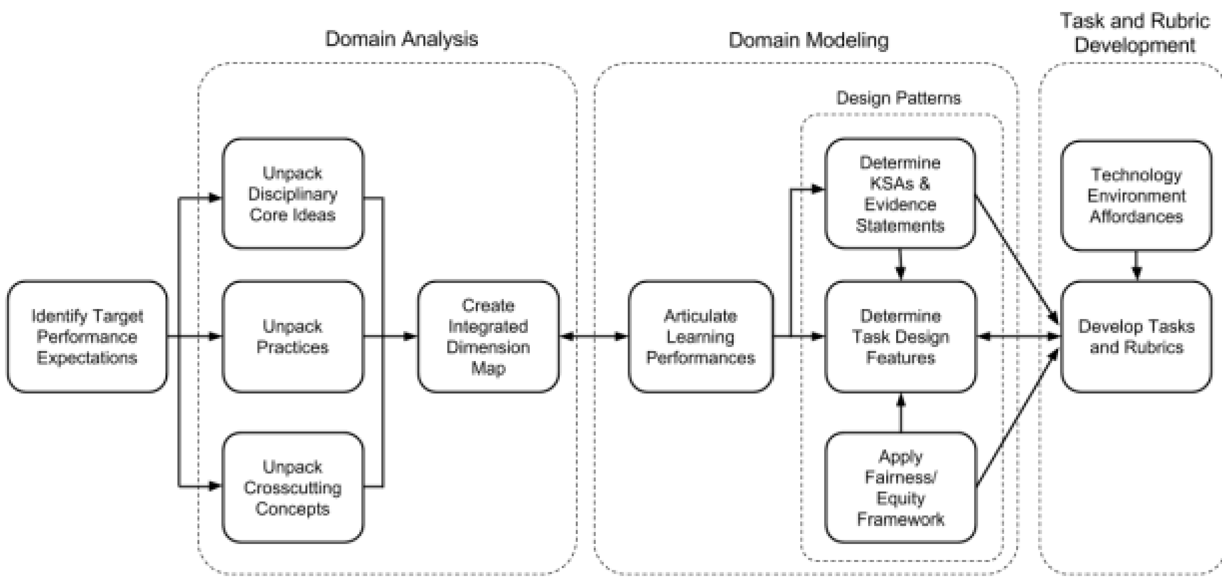


Figure 1. Assessment design process schematic.

Unpacking the NGSS dimensions. The disciplinary core ideas in the target PEs are unpacked by elaborating on and documenting the meaning of key terms, determining assessment boundaries for content knowledge, and identifying the background knowledge expected of students to develop grade-level-appropriate understanding of a disciplinary core idea. This elaboration extends what is in the *Framework* and NGSS by identifying research-based problematic student ideas. We unpack the science practices and crosscutting concepts by identifying and defining their core aspects, identifying intersections with other practices and crosscutting concepts, articulating the knowledge, skills, and abilities (KSAs) associated with them, and articulating the evidence that would demonstrate students possess and can use the KSAs.

Creating integrated dimension maps. We use the elaborations in the unpacking process to represent the conceptual terrain for achieving each target PE. Integrated dimensions maps describe the essential disciplinary relationships and link them to aspects of the targeted crosscutting concepts and science practices. The maps also illustrate how teachers can support students over time to meet each targeted PE. These maps are essential to the principled articulation of three-dimensional learning performances that coherently represent each target PE. To develop the integrated dimension maps, we begin by creating a traditional concept map (e.g., Schwendimann, 2015) of the essential aspects of disciplinary core ideas included in the target PE. We identify the key concepts and express their disciplinary relationships using labeled relationships between the concepts. Next, we add the aspects of the practices and crosscutting concepts that are best aligned with the disciplinary relationships for the purpose of creating learning performances and assessment tasks. The resulting dimension map expresses a range of ways that the three NGSS dimensions may be coherently integrated into learning performances, which collectively represent what students would be expected to know and be able to do, if they are proficient with the knowledge and skills underlying the target PE.

Domain Modeling

In ECD, domain modeling entails organize information from the Domain Analysis to link task and rubric design to evidence of student proficiency in the domain. We first articulate a set of learning performances based on the integrated dimension map. These learning performances represent a principled decomposition of the target PE and constitute the claims we wish to make about what students know and can do. We then create a design pattern (Mislevy & Haertel, 2006) for each learning performance that describes the design specifications for tasks and rubrics to be aligned with the learning performance.

Articulate learning performances. Using the integrated dimension maps, we articulate a set of learning performances that collectively represent a target PE. Learning performances integrate specific aspects of the dimensions comprising (or closely related to) the target PE, as identified in the Domain Analysis. Multiple learning performances are designed to build toward a PE in a way that could inform a teacher about a student's progress toward proficiency with the PE. Learning performances therefore not only represent the content of the PEs, but also address practical, intermediate targets for instruction aligned with the PEs.

Table 1 lists a set of four learning performances for PE MS-PS1-5, illustrating how learning performances identify specific, intermediate performance targets for instruction that guides students toward the PE. A set of learning performances can also vary in disciplinary complexity. For example, learning performance 4 addresses the disciplinary ideas of atom conservation and the production of new substances, while learning performance 7 addresses these ideas in addition to conservation of mass and the regrouping of atoms. We do not claim that these particular learning performances are the only ones that could be derived for PE MS-PS1-5. Specific decisions about learning performances should address the specific needs of teachers, learners, and alignment with curriculum materials, among other considerations.

Table 1. A set of four learning performances for NGSS PE MS-PS1-5. Each learning performance constitutes a claim about what students should be able to do as they make progress toward a PE.

LP 4	Students evaluate whether a model explains that a chemical reaction produces new substances and conserves atoms.
LP 5	Students evaluate whether a model explains that a chemical reaction produces new substances and conserves mass because atoms are conserved.
LP 6	Students use a model to explain that in a chemical reaction atoms are regrouped and why mass is conserved.
LP 7	Students develop a model of a chemical reaction to explain that new substances are formed by the regrouping of atoms and that mass is conserved.

Specify design patterns. For each learning performance, we specify a design pattern that guides the design of assessment tasks and rubrics that are aligned to it. Design patterns may include a wide range of task design specifications as part of a broader process of Domain Modeling (Mislevy & Haertel, 2006). These specifications may include task features derived via the application of a fairness and equity framework (Alozie, Fujii, Leones, Cheng, Pennock, & Damery, 2017), or variable task features that can shift the difficulty or focus of a task (Harris, et al., 2016). We focus on three types of design specifications that provide the basis for tightly integrating task and rubric design: (1) focal knowledge, skills, and abilities (KSAs), (2) features of student responses that constitute evidence of each focal KSA, and (3) characteristic features of assessment tasks that can effectively elicit this evidence of proficiency. We describe these specifications below and illustrate examples in Table 2. Figure 2 illustrates schematically the relationships among learning performances, assessment tasks, focal KSAs, evidence statements, and rubric components (further described below in the Assessment Task and Rubric Development section).

Table 2. Excerpts from the design pattern for learning performance 7, including focal KSAs, evidence statements, and characteristic task features.

Learning Performance	Students develop a model of a chemical reaction that explains new substances are formed by the regrouping of atoms, and that mass is conserved.
Focal Knowledge, Skills, and Abilities	• FKSA J: Ability to support model use, development, or evaluation by explaining that a chemical reaction conserves atoms and/or mass • FKSA L: Ability to support model use, development, or evaluation by explaining that chemical reactions regroup atoms • FKSA M: Ability to develop a model of a chemical reaction that regroups and conserves atoms
Evidence Statements	Student's response includes • Evidence for FKSA J: Student states how many of each type of atom are shown in the model before and after the process occurs. • Evidence for FKSA L: Student states that the atoms in the model are regrouped during the process. • Evidence for FKSA M: Students draws a model whose atoms are correctly regrouped from reactants to products and that conserves each type of atom.
Characteristic Task Features (selected)	Each task must include • A description of a scenario involving a simple chemical reaction • 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 explain how the model explains why mass is conserved in the reaction • A prompt to explain what happens to the atoms during the reaction

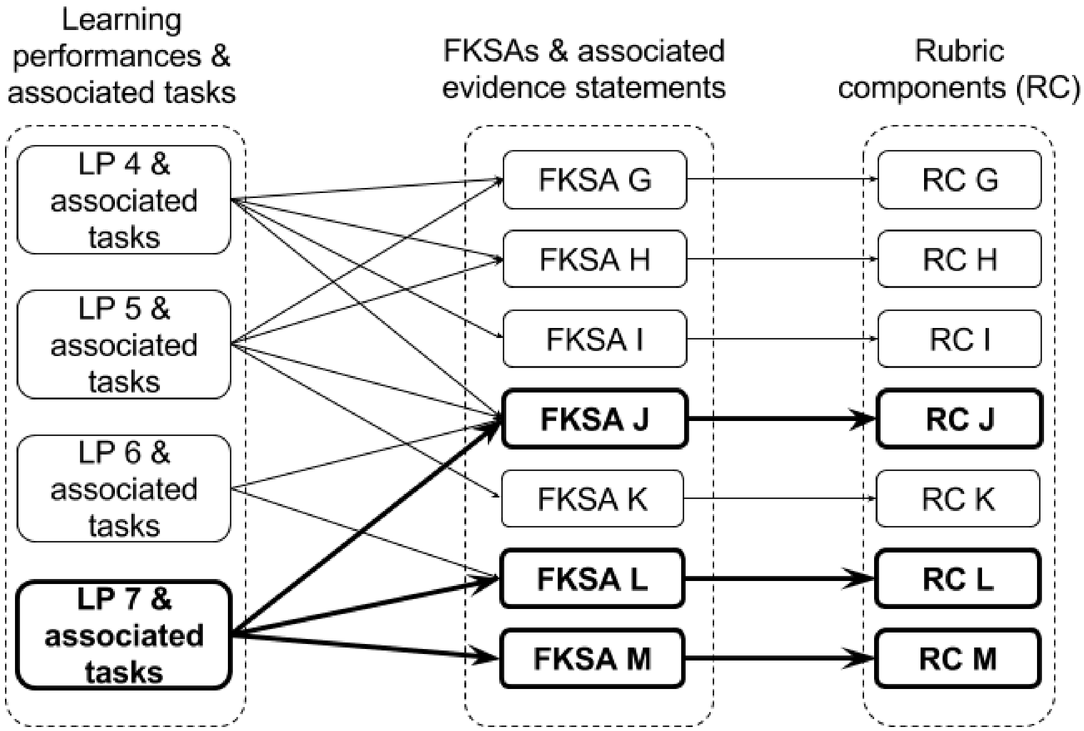


Figure 2. Schematic illustrating the relationship among learning performances, tasks, focal KSAs, and rubric components. The focal KSAs and rubric components associated with a particular learning performance (LP 7) are bolded.

Focal knowledge, skills, and abilities (FKSAs). Though we specify learning performances that can be assessed with a single task, learning performances integrate multiple and distinct aspects of the target proficiencies. For a set of learning performances associated with an NGSS PE, we identify several focal KSAs that represent distinct aspects of proficiency of interest to a teacher using one or more tasks for formative assessment. Each focal KSA integrates three NGSS dimensions associated with a learning performance, ensuring that the evidence of proficiency elicited by the tasks and the scoring approach for the tasks reflect the knowledge-in-use perspective, rather than isolating the dimensions from one another. Each task elicits evidence of proficiency with all FKSAs associated with its learning performance, and rubrics exhibit distinct components that correspond to the focal KSAs. Table 2 lists the three focal KSAs associated with learning performance 7. Each of these focal KSAs integrates disciplinary core ideas around chemical reactions, an aspect of the practice of developing models, and aspects of the crosscutting concept of energy and matter concerning matter conservation.

Because learning performances addressing the same PE necessarily exhibit some overlap in the claims they make about student proficiency, a specific focal KSA may correspond to multiple learning performances. For example, the ability to make a statement about a model of a chemical reaction constitutes an aspect of proficiency with using, developing, and evaluating models. Identifying focal KSAs in a way that expresses this overlap among learning performances allows evidence of these aspects of proficiency to be elicited across tasks and learning performances.

Evidence statements. Each focal KSAs constitutes the basis for a corresponding evidence statement. Evidence statements describe observable features of student performance that provide evidence of proficiency with a focal KSA. Evidence statements inform the specification of assessment task features and represent the highest scoring level for rubric components addressing a focal KSA. Table 2 lists the evidence statements corresponding to each focal KSAs associated with learning performance 7.

Assessment task features. Design patterns also describe features of assessment tasks that effectively elicit evidence of proficiency with the focal KSAs (Mislevy & Haertel, 2006). Characteristic features are task features that must be included in a particular type of task to ensure that the task can elicit the target proficiency. Some characteristic features are intended to minimize the need for construct irrelevant knowledge and skill for completing the task; others are derived from a fairness and equity framework to make the tasks accessible and fair to diverse students. Table 2 focuses on characteristic features, such as technology tools and prompts, that are needed to directly elicit evidence of the focal KSAs associated with learning performance 7.

Assessment task and rubric development

We use the design patterns to guide the development of both assessment tasks and their corresponding rubrics, ensuring that both tasks and rubrics align with a clear specification of the evidence to be derived from each student response.

Task development. We develop the tasks based on specifications in the design patterns, in order to elicit evidence of proficiency as described by the focal KSAs and evidence statements of the corresponding learning performance. In our approach, a task is not synonymous with a single prompt and its associated student response. Instead, it comprises a scenario that might include multiple prompts. A task might require various types of student responses including selections, written elements, drawings, or interactions with simulations. We intend for teachers to use specific tasks at appropriate points during instruction to gauge their students' progress toward achieving a PE. Figure 3 illustrates an example assessment task aligned with learning performance 7. The drawing tool students would use to illustrate the model is not pictured. The task aligns with the design specifications listed Table 2, affording students the opportunity to produce the evidence of proficiency described in the evidence statement. This task, by necessity, includes all of the characteristic features required to elicit the three FKSAs associated with the learning performance.

Battery in Tap Water: Rosy put a battery in a beaker of tap water. She observed gas bubbles coming from the positive and negative ends of the battery, as shown in the video below.

She tested the bubbles and found that some of the bubbles were made of hydrogen gas and some were made of oxygen gas.

Draw a model that shows the chemical reaction of water changing into hydrogen and oxygen gas.

Based on your model, describe
- what happens during the reaction to the atoms of the water molecules, and
- how your model explains why mass is conserved during this reaction.

Figure 3: Battery in Tap Water task, aligned with Learning Performance 7 and eliciting evidence of proficiency with FKSAs J, L, and M. The drawing tool associated with the task is not pictured.

Rubric development. The rubric development approach centers on the development of multiple rubric components, each of which corresponds to distinct aspects of proficiency of interest to teachers for classroom assessment. The rubrics are intended for use by researchers and aim to promote high scoring reliability. In the Discussion and Next Steps section (below), we discuss our preliminary efforts to create classroom rubric appropriate for use by teachers during the course of instruction.

We use the focal KSAs and evidence statements from the design patterns to develop a scoring rubric for each assessment task. In order to keep students' proficiencies with each focal

KSA distinct from one another, each rubric component measures proficiency with a specific focal KSA and builds from the corresponding evidence statement (Figure 2). Table 4 illustrates an example of a scoring rubric for an assessment tasks associated with the Battery in Tap Water task aligned with Learning Performance 7. When scoring a student's responses to a task, scorers would apply each rubric component to the response, obtaining a set of scores that collectively describe the student's proficiency with the learning performance. In addition to separating multiple aspects of a student's proficiency needed to respond correctly to a task, the individual rubric components focus scorers' attention on specific features of a student's responses, promoting reliability in scoring.

The approach has several advantages over using a typical holistic rubric to score student responses. First, multiple rubric components allow scoring decisions to be more tightly linked to distinct aspects of student proficiency. Second, rubric components allow scorers to focus on a small number of student response features at a time, without the need to consider multiple aspects of the student response simultaneously. Third, because each rubric component is aligned to a specific aspect of proficiency, they have the potential to inform teachers about what aspects of the PE their students need guidance with.

In preliminary reliability studies on the rubrics, scorers reported subjectively that applying individual rubric components one at a time to student responses helped them focus on specific features of students' responses. Making a series of separate scoring judgments appeared to reduce cognitive demands on scorers relative to assigning a single, holistic score based on multiple judgments. These reduced demands have the potential to limit scoring errors and improve reliability. In these preliminary studies, we observed that initial versions of our rubrics did not adequately express the key distinctions between the aspects of proficiency for scorers. Subsequent refinements to the rubrics clarified the fundamental conceptual distinctions between rubrics for the same task and helped scorers apply the rubrics reliably. We conjecture that in order for our rubric approach to be successful, the conceptual uniqueness of each FKSA must be articulated clearly, to avoid conflating the aspects of proficiency being assessed. Furthermore, we observe that a rigorous training process for scorers is essential to help them maintain distinctions between multiple scores on a single task.

Table 4. Example rubric for the Battery in Tap Water Task, aligned with learning performance 7. The rubric components address focal KSAs J, L, and M, as illustrated in Figure 3 above. Components J and M are scored polytomously from 0 to 2, while component L is scored dichotomously from 0 to 1.

Score	Rubric Component J	Rubric Component L	Rubric Component M
2	Student states • Both sides have 2 O atoms AND • Both sides have 4 H atoms		Student model shows • 2 oxygen atoms and 4 hydrogen atoms on each side AND • H_2O as the reactant and H_2 and O_2 as the products
1	Student states • Both sides have 2 O atoms or 4 H atoms OR • Same number of H and O atoms on each side 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	Student model shows • Equal numbers of O and H atoms on each side OR • H_2O as the reactant and H_2 and O_2 as the products
0	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	Student model includes • Unequal numbers of O and H atoms on each side AND • Reactants/products other than H_2O, H_2, or O_2

Discussion and Next Steps

Our integrated design approach for assessment tasks and rubrics promotes coherence across the three-dimensional nature of NGSS PEs, assessment task design features, rubric design, evidence-based scoring, and classroom-based formative assessment. Reflecting the knowledge-in-use perspective, each design phase maintains the integration of all three NGSS dimensions. A central question we have wrestled with is whether rubrics should integrate the dimensions into a single, holistic score or separately evaluate performance on each of the three dimensions. We chose to take an integrated approach to interpreting student responses, thereby staying true to the *Framework and NGSS* vision of science proficiency as the ability integrate the three dimensions.

The breadth and complexity of the end-of-grade-band NGSS PEs requires teachers to formatively diagnose specific aspects of proficiency with the PE in order to support instruction. Our ECD-based approach to task and rubric design identifies these aspects of proficiency and provides a framework for articulating a scoring approach that can help teachers guide their students toward achieving these complex performances. Using rubrics having distinct components that distinguish key aspects of proficiency therefore represents an important advance in aligning instruction and assessment aligned with NGSS. Our work shows that rubrics that can distinguish key aspects of proficiency while maintaining the integration of the three dimensions appear to be feasible. The scoring approach has the potential to provide new insights into students' science proficiency and its change over time with instruction.

Because rubrics such as the one shown in Table 4 are articulated for research purposes, we are currently examining how to express the rubrics in a form that is both practical for classroom use and educative for teachers about the nature of the NGSS. From a rubric as illustrated above in Table 4, 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 as guiding questions in order to highlight for teachers how to evaluate students' proficiency with integrating the three NGSS dimensions. We include a preliminary example of a classroom rubric for the Battery in Tap Water task in the Appendix. Designing the classroom rubrics to be both useful and educative for classroom teachers suggests the need for the tasks and rubrics to be reviewed and tested by expert school and/or district practitioners and to be refined on the basis of these studies to improve their utility.

Finally, though our design approach is specifically tailored to the NGSS, we believe the approach could be used to develop assessments of any multidimensional performance constructs that integrate content knowledge and disciplinary practices. For example, the Framework for K-12 Computer Science Education (www.k12cs.org) identifies computing concepts and practices as dimensions of proficiency with computer science. Our approach could help assessment developers address integrated performance statements aligned with the computer science framework.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants DRL-1316903, -1316908, and -1316874 and by the Gordon and Betty Moore Foundation under Grant 4482. Any opinions, findings, and conclusions or recommendations expressed in this document are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Gordon and Betty Moore Foundation. We also greatly appreciate the contributions of Christopher Harris to this manuscript.

References

Alozie, N., Fujii, R., Leones, T., Cheng, B., Pennock, P. H. & Damery, K. (2017, May). An equity framework for the design and development of NGSS aligned formative assessment tasks. Paper presented in a poster session at the American Educational Research Association Annual Meeting, San Antonio, TX.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

DeBarger, A. H., Harris, C. J., D'Angelo, C., Krajcik, J., Dahsah, C., Lee, J., & Beauvineau, Y. (2014). Constructing assessment items that blend core ideas and science practices. In Polman, J. L., Kyza, E. A., O'Neill, D. K., Tabak, I., Penuel, W. R., Jurow, A. S., O'Connor, K., Lee, T., and D'Amico, L. (Eds.). *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014, Volume 3*. Boulder, CO: International Society of the Learning Sciences.

Duschl, R. & Osborne, J. (2002). Argumentation and discourse processes in science education. *Studies in Science Education*, *38*, 39-72.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K.W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications*. Menlo Park, CA: SRI International.

Harris, C. J., McNeill, K. L., Lizotte, D. L., Marx, R. W., & Krajcik, J. (2006). Usable assessments for teaching science content and inquiry standards. In M. McMahon, P. Simmons, R. Sommers, D. DeBaets & F. Crowley (Eds.), *Assessment in science: Practical experiences and education research* (pp. 67-88). Arlington, VA: National Science Teachers Association Press.

Harris, C. J., & Salinas, I. (2009). Authentic science learning in primary and secondary classrooms. In M. I. Saleh & M.S. Khine (Eds.), *Fostering scientific habits of mind: Pedagogical knowledge and best practices in science education* (pp. 125–144). Netherlands: Sense.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta- analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28-37.

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

Latour, B., & Woolgar, S. (1996). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy: Supporting development in learning in contexts. In W. Damon, R. M. Lerner, K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology, 6th ed.* (Vol. 4). Hoboken, NJ: John Wiley and Sons.

McElhaney, K. W., Gane, B. D., Harris, C. J., Pellegrino, J. W., DiBello, L. V., & Krajcik, J. S (2016, April). *Using learning performances to design three-dimensional assessments of science proficiency*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

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

National Research Council. (2000). *How people learn: Brain, mind, experience*. (Expanded ed.). Washington, DC: The National Academies Press.

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. (2009). *Learning science in informal environments: People, places, and pursuits*. 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: National Academies Press.

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

Pellegrino, J. W., Wilson, M., Koenig, J. & Beatty, A. (Eds.) (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: The National Academies 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*(1), 57-84.

Schwendimann, B. A. (2015). Concept maps as versatile tools to integrate complex ideas: From kindergarten to higher and professional education. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, *7*(1), 73-99.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, *46*(6), 610–631.

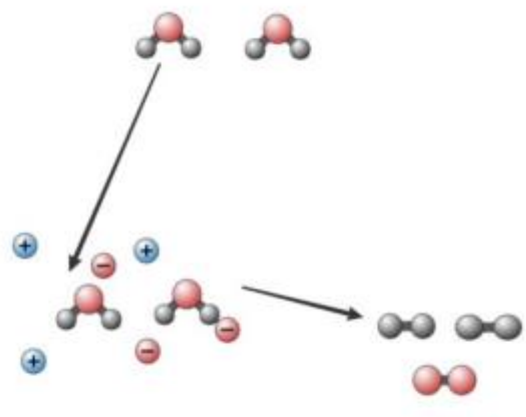
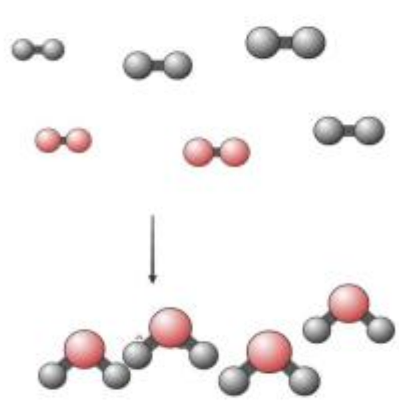
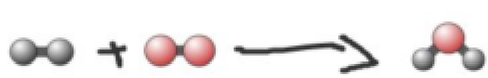
Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, *34*, 254–284.

Appendix: Preliminary example of a classroom rubric

<table>
<tr><td colspan="2">Component J: Develop a model of matter conservation in a reaction by atom conservation and atom rearrangement.</td></tr>
<tr><td>Evidence of proficiency in student work</td><td>Exemplar student responses at each level</td></tr>
<tr><td>Does the student model show matter conservation in a reaction by showing the (A) the correct reactant and product molecules AND (B) that the reaction conserve atoms?

☐ Both A and B (high)

☐ Either A or B (partial)

☐ Neither A nor B (low)</td><td>
(high)

(partial)

(low)</td></tr>
</table>

Component L: Use a model to explain matter conservation in a reaction by atom conservation.	
Evidence of proficiency in student work	**Exemplar student responses at each level**
Does the student describe that mass is conserved because the model shows that atoms are conserved with 4 oxygen and 4 hydrogen atoms on each side of the reaction? ☐ Yes (high) ☐ States either oxygen or hydrogen are the same (partial) ☐ States atoms are conserved without stating numbers (partial) ☐ No (low)	The mass is conserved in the model because there are four oxygen atoms and four hydrogen atoms on both sides. (high) The model shows the same number of atoms on each side of the reaction. (partial) The atoms begin to separate in the model. Mass is conserved during this reaction because it is a chemical reaction. (low)

Component M: Use a model to explain matter conservation in a reaction by atom rearrangement.	
Evidence of proficiency in student work	**Exemplar student responses at each level**
Does the student describe that a chemical reaction occurred because the same atoms are rearranged (e.g., reactants broken apart and products are formed) in the model? ☐ Yes (high) ☐ No (low)	The oxygen molecules come apart and split into atoms and then one of the atoms attach to the two hydrogen atoms and together they make water (high) The atoms begin to separate. Mass is conserved during this reaction because it is a chemical reaction. (low)