Measuring what matters: Challenges and opportunities in assessing science proficiency

James W Pellegrino
The University of Illinois at Chicago

James W Pellegrino is Liberal Arts and Sciences Distinguished Professor and Distinguished Professor of Education at the University of Illinois at Chicago. He also serves as Co-director of UIC's interdisciplinary Learning Sciences Research Institute. His research and development interests focus on children's and adults' thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 300 books, chapters and articles in the areas of cognition, instruction and assessment. His current research is funded by the National Science Foundation and the Institute of Education Sciences. He has served as head of several National Academy of Sciences study committees, including co-chair of the Committee on Learning Research and Educational Practice, and co-chair of the Committee on the Foundations of Assessment which issued the report *Knowing What Students Know: The Science and Design of Educational Assessment*. Most recently he served as a member of the Committee on Science Learning: Games, Simulations and Education, as chair of the Committee on Defining Deeper Learning and 21st Century Skills, and co-chair of the Committee on Developing Assessments of Science Proficiency in K–12. He is a past member of the Board on Testing and Assessment of the National Research Council and a lifetime member of the National Academy of Education. He currently serves on the Technical Advisory Committees of several states, agencies and consortia of states funded under the United States' Department of Education Race to the Top assessment initiative.

Abstract

A key challenge in shaping science learning for the future will be to develop new measures of learning that take into account what it means to be proficient in science (Pellegrino, 2013). The emergent view on proficiency, grounded in learning sciences research, emphasises using and applying knowledge in the context of disciplinary practice. Referred to as knowledge-in-use, this perspective on science proficiency is a centrepiece of the United States' National Research Council's (NRC) *Framework for K–12 Science Education* (NRC, 2012), embodied in the new US national standards (NGSS Lead States, 2013) and emphasised in the recently released NRC report on developing assessments to measure science proficiency (Pellegrino, Wilson, Koenig & Beatty, 2014). Central to this view is that disciplinary content — both disciplinary core ideas and crosscutting concepts — and practice should be integrated. This would mean that as students apply knowledge to make sense of phenomena and solve problems, they deepen their conceptual understanding of content as well as their understanding of how to do science. This paper provides a brief overview of a systematic and scalable approach for designing assessment items to measure student proficiency with new science learning goals that blend disciplinary core ideas and crosscutting concepts with practices. The assessment tasks are intended for formative use within classroom instruction. Drawing on prior research from assessment and curriculum design (for example, see DeBarger, Krajcik & Harris, 2014; DeBarger, Penuel & Harris, 2015), this paper presents such a design approach and considers implications of the overall work in this field.

Conceptual framework

The prior generation of US science standards (for example, NRC, 1996, 2000) treated content and inquiry as fairly separate strands of science learning, and assessments followed suit. In some respects, the form the standards took contributed to this separation: content standards stated what students should know, and inquiry standards stated what they should be able to do. Consequently, assessments separately measured the knowledge and practice components. The shift to integrating science practices with disciplinary core ideas and crosscutting concepts, as emphasised in new US standards, called the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013), is based upon studies of actual scientific practice and what we currently know about student learning (cf., recent synthesis reports such as *Taking Science to School*, NRC, 2007 and *A Framework for K–12 Science Education*, NRC, 2012). This research corpus points to the importance of integrating content (that is, disciplinary core ideas and crosscutting concepts) and practice by emphasising that rich science learning requires tight coupling of what students know and what they can do. This idea of science *performance* (NGSS Lead States, 2013) presents a different way of thinking about science proficiency in that disciplinary core ideas and crosscutting concepts serve as thinking tools that work together with scientific and engineering practices to enable learners to solve problems, reason with evidence, and make sense of phenomena (NRC, 2012). The idea of science performance also signifies that measuring proficiency solely as acquisition of core content knowledge is no longer sufficient.

Knowledge-in-use learning goals comprise the Next Generation Science Standards and are articulated as *performance expectations*. Each performance expectation combines a science or engineering practice, disciplinary core idea, and 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. This integrated, knowledge-in-use perspective poses challenges for assessment design. At this time, there are very few examples of assessments that integrate science content and practices in a manner consistent with a knowledge-in-use perspective. There is tremendous need for this assessment design work, as assessment will play a central role in supporting implementation of the new directions in science education both in the US and internationally. Our approach to meeting this challenge uses principles of evidence-centred design (Almond, Steinberg & Mislevy, 2002). Evidence-centred design has been used in wide-ranging assessment design contexts, from the development of large-scale, high-stakes assessments to the design of classroom-based assessments and other proximal or close measurement instruments. Evidence-centred design emphasises the evidentiary base for specifying coherent, logical relationships among: (1) the learning goals that comprise the constructs to be measured (that is, the claims we want to make about what students know and can do); (2) the evidence in the form of observations, behaviours or performances that should reveal the target constructs; and (3) the features of tasks or situations that should elicit those behaviours or performances. The need for a principled approach to assessment design, such as evidence-centred design, was explicitly discussed in the United States' National Research Council report on developing assessments aligned to the Next Generation Science Standards (Pellegrino et al., 2014).

Application of evidence-centred design to three-dimensional science assessment

Figure 1 provides an overview of our overall design process for constructing assessment tasks that align with the Next Generation Science Standards. Our process follows the logic of evidence-centred design and contains three distinct phases — unpacking (domain analysis), constructing an assessment argument (domain modelling), and task and rubric development. While the figure represents a linear process that begins with selecting performance expectations and unpacking the three dimensions, it is important to realise that the process is very iterative in nature. The step of specifying evidence statements, for example, has caused us to revisit and revise our learning performances and unpacking.

Domain analysis: Unpacking components of performance expectations

In evidence-centred design, domain analysis typically entails gathering substantive information about how knowledge is acquired and used in a domain such as physical or life science. *A Framework for K–12 Science Education* and the *Next Generation Science Standards* specify meaningful ways to integrate the content and practices to promote assessment of learning in each domain. The analyses of the domain inform the construction of learning performances that represent formative assessment opportunities to check in on student progress toward performance expectations.

Unpacking the disciplinary core ideas

In this phase of evidence-centred design, we first unpack core ideas associated with a cluster of *Next Generation Science Standards* performance expectations at a given grade level or grade band by elaborating the meaning of key terms, defining expectations for understandings for the targeted student level, determining assessment boundaries for content knowledge; identifying background knowledge that is expected of students to develop a grade-level-appropriate understanding of a disciplinary core idea; and considering research-based problematic student ideas and misconceptions.

Unpacking the science practices

Our unpacking of the science practices involves consideration of the core components of the practice, intersections with other science practices and the evidence required to demonstrate the practice.

Unpacking the crosscutting concepts

Unpacking the crosscutting concepts involves identifying the important components and opportunities for intersections with the science practices and with the particular disciplinary core ideas that are the target of the assessment.

Domain modelling: Specifying a knowledge-in-use assessment argument

Leveraging the unpacking of science practices, crosscutting concepts and disciplinary core ideas, we then move toward specifying a knowledge-in-use assessment argument. In this step, 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 (see Table 1). Our claims, evidence, and task features reflect a knowledge-in-use perspective in that we emphasise the application of core ideas and crosscutting concepts through engagement in a science practice. Each claim takes the form of what we refer to as a *learning performance*. Each learning performance clearly describes what we expect students to demonstrate to provide evidence that they have achieved an aspect of a performance expectation. To construct learning performances, we identify the key aspect(s) of a disciplinary core idea, practice and crosscutting concept from our unpacking work, to specify statements of what a student should be able to do. As such, learning performances integrate *aspects* of disciplinary core ideas, practices and crosscutting concepts, and are written to express knowledge in use. Learning performances, however, are of a smaller grain size than performance expectations. Together, a set of learning performances provides the detail needed to create a coherent and bundled set of assessment tasks that would provide evidence that students can use the knowledge aligned to a performance expectation or cluster of performance expectations. In this way, high-quality learning performances function in relation to other learning performances to identify 'what it takes' to make progress toward meeting a standard (for example, *Next Generartion Science Standards* performance expectations). Learning performances are also helpful for teachers as they help to identify an important opportunity that teachers should attend to and assess *before* the end of an instructional unit.

Once a learning performance has been specified, we then express the evidence students need to demonstrate to show they have met the claim. This can be thought of as student behaviours or performances that provide evidence of attaining the learning performance. To complete our assessment argument and before we can write assessment tasks, we need to specify characteristic and variable task features. Characteristic task features describe the attributes that are common across all the tasks for a learning performance. For instance, one characteristic task feature is that all tasks need to provide a motivating context. Variable task features describe what features can vary across the tasks. For instance, the level of scaffolding is one example of a variable task feature. Table 1 presents a knowledge-in-use assessment argument for a claim integrating disciplinary content knowledge about structure and properties of matter and the crosscutting concept of patterns with scientific practice of constructing a scientific explanation.

Developing tasks and rubrics

The final phase of the design process involves using the information detailed in the assessment argument to develop actual assessment tasks that will be presented to students. The task design depends on the specification of the characteristics and variable task features and allows for assembly of multiple tasks within a 'family' where the variations among the tasks could readily reflect intended levels of challenge. The task design process also takes into account the forms of evidence needed to support the learning performance claim and the ways in which that evidence will be scored and evaluated for purposes of rubric development. Obviously, validation of our assumptions about the tasks depends on collecting various forms of empirical data from students under conditions where we have a reasonable set of assumptions of the prior opportunity to learn.

Discussion and implications

Our design approach provides a broadly accessible vision of how to design Next Generation Science Standards assessments and is a vehicle for documenting principled design decisions. The systematic process anchored in evidence-centred design allows us to create well-aligned tasks that are usable across varied classroom environments. Although we have focused our efforts to date on physical sciences disciplinary core ideas and only a subset of the scientific and engineering practices, our process should generalise to other core ideas, crosscutting concepts and practices.

Figure 1: Design process for developing assessment items aligned to the *Next Generation Science Standards*

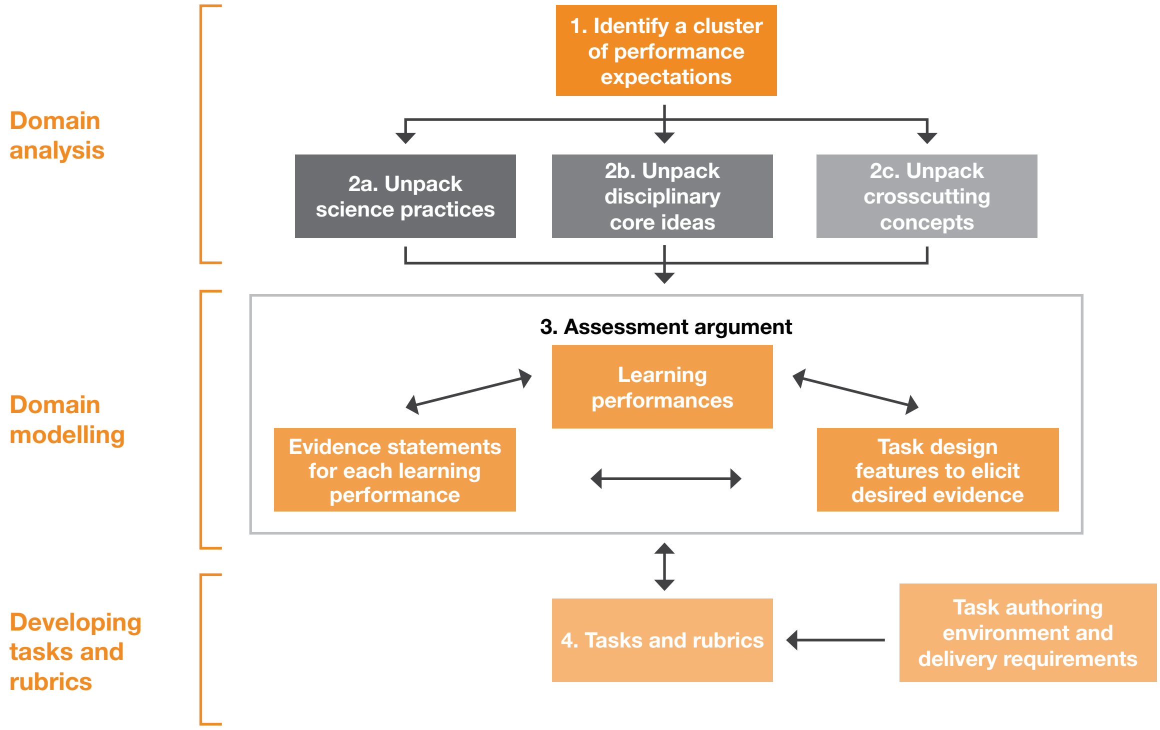

Table 1 Knowledge-in-use assessment argument

Learning performance (claim)	• Students should be able to construct an explanation about how they determine substances are the same based upon characteristic properties
Additional knowledge, skills and abilities	• Knowledge that some properties can be used to identify substances, and that these properties are called characteristic properties (e.g., density, melting point, boiling point) • Knowledge that temperature, volume, and mass cannot be used to identify substances and are not characteristic properties • Ability to identify patterns in data on physical properties of different substances • Ability to identify which data can be used as valid and appropriate evidence • Knowledge that a scientific explanation includes a claim, evidence and reasoning
Evidence required to demonstrate proficiency	• Written claim: statement that substances (e.g., Liquid A and B) are the same or are different • Stated evidence: identification of at least two characteristic properties to support claim • Description of reasoning: statement that the same substance must have the same set of characteristic properties or that different substances have different characteristic properties
Characteristic task features	• Assessment is limited to analysis of the following characteristic properties: density, melting point, boiling point, solubility, flammability and odour • The term 'substance' means a pure material (not a mixture of substances). • Tasks provide data about characteristic properties of substances • Tasks provide a motivating/authentic context
Variable task features	• Types of properties included as data/evidence • State of matter of substances (i.e., solid, liquid, or gas state) • Inclusion of irrelevant data (e.g., non-characteristic properties) • Level of scaffolding to develop claim, evidence and reasoning

While our design approach has important advantages, challenges also exist. From a learning perspective, integrated assessment of key aspects of all three dimensions seems to be feasible and should provide insights into student achievement and its change over time with instruction. However, such an approach brings unique challenges from the perspective of measurement and interpretation of performance. A central question is whether rubrics should integrate the *Next Generation Science Standards* dimensions into a single score or separately evaluate aspects of performance for all the three dimensions. This involves issues related to ease of use and feasibility, including the extent to which each of the three performance components are separable and identifiable. Teachers will also need professional development on how to use these items in the classroom. Thus, creating models of how three-dimensional items can be used formatively in the classroom will be instrumental for effective classroom use.

We believe that our program of research and development will help to provide answers to critical questions related to the design and use of assessments of science knowledge in use. A critical need exists for research and development of high-quality assessments that align with the *Next Generation Science Standards* that express knowledge-in-use learning goals. More important, teachers need to be able to use these tasks in classrooms to provide themselves and students with information about progress towards meeting the performance expectations. Having exemplary formative assessments that integrate core disciplinary ideas, scientific and engineering practices and crosscutting concepts will be important to multiple stakeholders. Teachers, students, parents and school officials are interested in using high-quality assessments that provide information preparing students for university and career readiness in the fields of science, technology, engineering and maths. Assessment researchers need to better understand the design principles and psychometric properties of assessments that integrate core ideas, crosscutting concepts and science practices. Science education researchers want to use the assessments to better understand larger issues that widespread adoption of a three-dimensional learning perspective would entail, including developing and evaluating new science curricula. Science educators and policy–makers want assessments that help them to better understand students' knowledge and abilities and also to inform changes in classroom instruction.

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(5). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1671

DeBarger, A.H., Krajcik, J.S. & Harris, C.J. (2014, April). *Using the Framework to design assessments of modeling in physical and Earth science*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Pittsburgh, PA.

DeBarger, A.H., Penuel, W.R., Harris, C.J. & Kennedy, C.A. (2015, May). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, OnlineFirst. doi:10.1177/1098214015581707

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

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

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

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

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

Acknowledgements

Project members contributing to this work include Chanyah Dahsah, Dan Damelin, Cynthia D'Angelo, Louis DiBello, Brian Gane, Christopher Harris, Joseph Krajcik, Jane Lee, and Kevin McElhaney. We also gratefully acknowledge Angela DeBarger from Lucas Education Research, Deborah Peek-Brown and Tali Tal from Michigan State University, Gauri Vaishampayan from University of Illinois at Chicago, Frieda Reichsman from Concord Consortium, and Tiffany Leones and Reina Fujii from SRI International for their contributions to this research and development project. This paper is based upon work supported by the National Science Foundation under Grant Numbers DRL- 1316903, 1316908, 1316874. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.