Designing NGSS Assessments to Evaluate the Efficacy of Curriculum Interventions

Angela Haydel DeBarger, William R. Penuel,
and Christopher J. Harris

September 2013

K-12 Center at ETS

Invitational Research Symposium on
Science Assessment

K-12 Center at ETS
Invitational Research Symposium on Science Assessment

Designing NGSS Assessments to Evaluate the Efficacy of Curriculum Interventions

Angela Haydel DeBarger
SRI International

William R. Penuel
University of Colorado Boulder

Christopher J. Harris
SRI International

Introduction

The *Framework for K-12 Science Education* (National Research Council [NRC], 2012) presents an ambitious vision for improving teaching and learning in science, and the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013) embody a key part of that vision by specifying challenging performance expectations that all students are expected to meet. Realizing the vision of the Framework, though, will require more than state adoption of the standards. It will require changes across the educational system to support implementation, including changes to curriculum, instruction, teacher professional development, and assessment (NRC, 2012).

With respect to assessment, the changes required will depend on the purposes for which assessments are intended to be used. Program evaluation is one of three different purposes for assessment identified in the Framework. Programs that are the object of evaluation in science education include curricula, in-service professional development programs, and focused interventions that target specific populations of students, either in or out of school. Since the Sputnik era, policy makers have provided funding to develop and test these kinds of programs in science education, on the premise that such programs provide necessary support for teachers to meet new and ambitious goals for student learning. Evaluations of these programs require assessments that are aligned to standards, in order to test this core premise.

Some of the challenges of developing assessments of NGSS for program evaluation are similar to challenges of developing assessments for other (e.g., formative, summative) purposes, but some are distinctive. For example, developers will need to construct and test tasks that elicit evidence related to

student learning. At the same time, the tasks should do more than elicit students' declarative knowledge. They must elicit evidence related to students' integration of knowledge of disciplinary core ideas, engagement with scientific practices, and facility with building connections across ideas (crosscutting concepts; NRC, 2012 Pellegrino, 2013). As with all assessments, developers must devise strategies for interpreting student responses to tasks, including developing reliable rubrics for tasks and identifying appropriate methods for statistical modeling of student scores. But designers must make choices about which groups or "bundles" of performance expectations to assess in a test, and they must decide whether and how to integrate disciplinary core ideas, science practices, and crosscutting concepts within rubrics.

Evaluation uses of assessment information present designers with additional challenges. Evaluators must choose assessments that are instructionally sensitive but that also permit fair comparisons among different treatments of the same performance expectations. Evaluators must employ research designs that generate compelling evidence related to the worth or value of programs, of which assessment data is one source among many. Evaluators consider evidence about implementation and the settings of implementation to be critical components of program evaluation. If programs are not implemented with integrity to principles of the program's design, inferences about the potential efficacy of the program may not be valid.

In this paper, we articulate an approach to designing assessments for the purpose of evaluating curriculum interventions intended to support implementation of NGSS. We outline the structure of a validity argument for such uses within an evidence-centered design (ECD; Mislevy & Haertel, 2006) framework. We then illustrate our approach by describing the design and testing of an assessment being used to evaluate middle school science curriculum materials in a cluster randomized trial in a large urban district. We highlight the challenges we faced, the design decisions we made, and rationale for those decisions. We conclude with recommendations for developing NGSS-aligned assessments for program evaluation.

Evaluation Context

The context for our assessment design work is an efficacy study of a commercially available middle school science curriculum that uses project-based science units to help students learn. Project-Based Inquiry Science (PBIS) is a comprehensive 3-year curriculum that is sold and distributed through It's About Time Publishing. Most of its units were developed in the context of a set of learning sciences research projects (Kolodner et al., 2008), notably the LeTUS program (Singer, Marx, Krajcik, & Clay-Chambers, 2000) and the Learning by Design project (Kolodner et al., 2003). In 2005, the developers began working closely with the publishing company to bring the curriculum to publication. The full curriculum became available to school districts during the 2009–2010 academic year.

The curriculum comprises science units in life, physical, and Earth science, spanning grades 6 through 8. A typical unit takes 8–10 weeks for teachers and students to complete. In contrast to some

other materials that present ready-made investigations for students to carry out, PBIS presents challenges to students in which they must investigate phenomena and apply concepts to answer a driving question or to achieve a design challenge. The driving question or challenge typically targets a core idea in science, and the activities within each unit provide students with multiple occasions for investigating as scientists would—through observations, asking questions, designing and carrying out experiments, building and using models, reading about the science they are investigating, constructing explanations, and so forth. In this way, the PBIS curriculum's design emphasizes a knowledge-in-use perspective (NRC, 2007) and reflects where the science education field is headed—teaching a few core ideas and integrating science practices.

The PBIS units that are the focus of our evaluation are in the areas of physical science (energy) and Earth science (processes that shape Earth's surface). The units address science ideas that align well to the core ideas articulated in the Framework and provide us with felicitous conditions for examining student learning of content integrated with two science practices named in the Framework and encompassed in NGSS performance expectations*:* developing and using models and constructing explanations.

At the time that we began our preparations for the evaluation study, available assessments did not integrate core ideas and science practices in the manner intended by the Framework and NGSS. We were faced, then, with needing to develop our own measures that had potential to elicit complex science practices blended with ambitious science content. We looked to existing models of assessment design for insight into approaches that might be relevant for guiding assessment design that aimed to systematically attend to this integration. As our effort was just getting underway, we were aware of other projects that were taking on the challenge of measuring the intersection of content and practice with a learning progressions perspective. The Learning Progressions in Middle School Science Instruction and Assessment (LPS) project, for instance, was examining learning progressions in physical science and in argumentation and scientific reasoning. Wilson (2009) described a unified approach to integrating science assessment with instruction that makes use of progress variables, typically multiple unidimensional Rasch scales that represent progress variables. The methods and findings of Wilson (2009) and of Briggs et al. (2006) were especially informative as we began to take on our own challenge of developing valid assessments that explicitly connected science content and practices as set forth in the Framework.

What we aimed to accomplish in our work was to systematically attend to the blending of content, practices and crosscutting concepts to design NGSS-aligned assessment tasks that students in the PBIS curriculum would have an opportunity to learn. We decided to use the ECD framework as a central strategy to articulate an assessment argument that would persuasively unpack performance expectations into a coherent association of learning goals, describe the kinds of tasks and situations that would elicit those goals, and demonstrate how particular performances can be interpreted as evidence for students' capabilities (Mislevy & Haertel, 2006).

Conceptual Framework

Assessment is a form of reasoning from evidence in which people use observations of students' actions and artifacts to make decisions and to support conclusions about what students know and can do (Pellegrino, Chudowsky, & Glaser, 2001). Validation of assessments entails developing a coherent, compelling argument that supports the conclusions or decisions, and such an argument requires many sources of evidence (Kane, 1992). Common sources of evidence used in validity arguments include expert judgments that particular items are aligned to standards, cognitive interviews, documentation of adequate levels of interrater reliability for scoring rubrics, cognitive interviews, and results of item response models. Evidence related to student opportunity to learn, the testing situation, and the plausibility of alternative conclusions may also be needed (Kane, 1992; Mislevy, 2007; Shepard, 1993). The type of evidence varies from situation to situation, depending upon the intended use of assessment information.

One important intended use of assessment information is to evaluate a curricular intervention. In a resource-constrained environment, policy makers and educational leaders need information about under what circumstances programs work and for whom they work (Means & Penuel, 2005). Such information can inform decisions about curricula and interventions to adopt, as well as decisions about how to support implementation and which students to target. The validity of conclusions about the worth of programs depends in part on the appropriateness of the student assessments used as measures of learning outcomes. Appropriate measures allow one to draw conclusions about what individual students can do, because the measures provide evidence regarding their utility and capacity to evaluate a particular intervention or set of programs.

A key initial goal for all assessment design is to design an assessment that allows for accurate judgments about what individual students know and can do. Assessments of student proficiency in meeting the performance expectations of the NGSS should generate evidence for claims about students' capacity to demonstrate understanding of core ideas and fundamental crosscutting concepts through engaging in science practices (Pellegrino, 2013). Developing validity evidence for such assessments presents a number of challenges, including specifying claims about what integrated knowledge of big ideas, practices, and crosscutting concepts looks like; identifying task features that could provide opportunities for students to demonstrate this integrated knowledge; and specifying the evidence from student responses to tasks that can support claims about their learning (Pellegrino, 2013). Developing a fair assessment is also challenging because few students today have had an adequate opportunity to learn according to the vision of the Framework. Moreover, students from nondominant communities historically have had lower access to the kinds of ambitious instructional practices called for in the Framework.

The ability to draw valid inferences about individual student learning from curriculum is not the only consideration in evaluating the worth or merit of programs, which is a key purpose of summative evaluation studies. Evaluators need to also consider whether their measures allow for fair comparisons

of a target program to compelling alternative programs (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). In addition, data on implementation are needed, because when programs are not implemented well, inferences about the potential efficacy of a program may not be valid (Love, 2004). Data on implementation and professional development are also useful for developing and testing hypotheses about the conditions under which programs can work (Confrey, Castro-Filho, & Wilhelm, 2000; Desimone, 2009; Means & Penuel, 2005).

Given the challenges and complexity of designing assessments for use in evaluation, it is useful to rely on tools for articulating a validity argument. Mislevy (2003, 2008) has argued that organizing a validity argument into a structure can help assessment designers plan validity studies that are needed to generate data, warrants for those data, and backing for warrants. The particular representations of assessment arguments used by Mislevy, which are adapted from Toulmin (1958), depict the kinds of validity evidence needed to justify conclusions about individual students, when the intended use is for placement or certification of individuals. Here, we represent a basic structure for organizing validity evidence for an assessment when the intended use is evaluating programs.

We represent this structure as a set of interlocking arguments, as shown in Figure 1. At the center, connecting the two arguments, is a claim or conclusion about what a student knows and can do. As they do for any assessment of student learning (Pellegrino et al., 2001), assessors need data on student actions, such as their response to a task or problem presented to them, in order to judge the strength of the claim. But whether the data support the claim depends upon a number of warrants that must be backed by other kinds of evidence: Was the task or problem adequate to elicit what students know and can do? Can student responses be reliably categorized or scored? Can students with knowledge in the domain understand the task directions and perform well on it? Do they do better than students with less understanding? Data concerning the assessment situation, too, are needed, such as the length of time students had to complete the assessment, and the level and kind of stress they might be under to perform well.

An important source of evidence related to claims about student proficiency is evidence of what Messick (1989) termed the *external component* of validity. This could include evidence that individuals' test scores are more strongly correlated with scores on other measures of the same construct than with scores on different constructs. It might also include evidence that measures are sensitive to the effects of instruction. Importantly, the comparisons made should be grounded in a theory of a construct; otherwise, correlations of scores on the measure being developed with other measures are difficult to interpret. Evidence of the external component of validity is likely to qualify claims about individuals, because "no single test is a pure exemplar of the construct but contains variants due to other constructs and method contaminants" (Messick, 1989, p. 48).

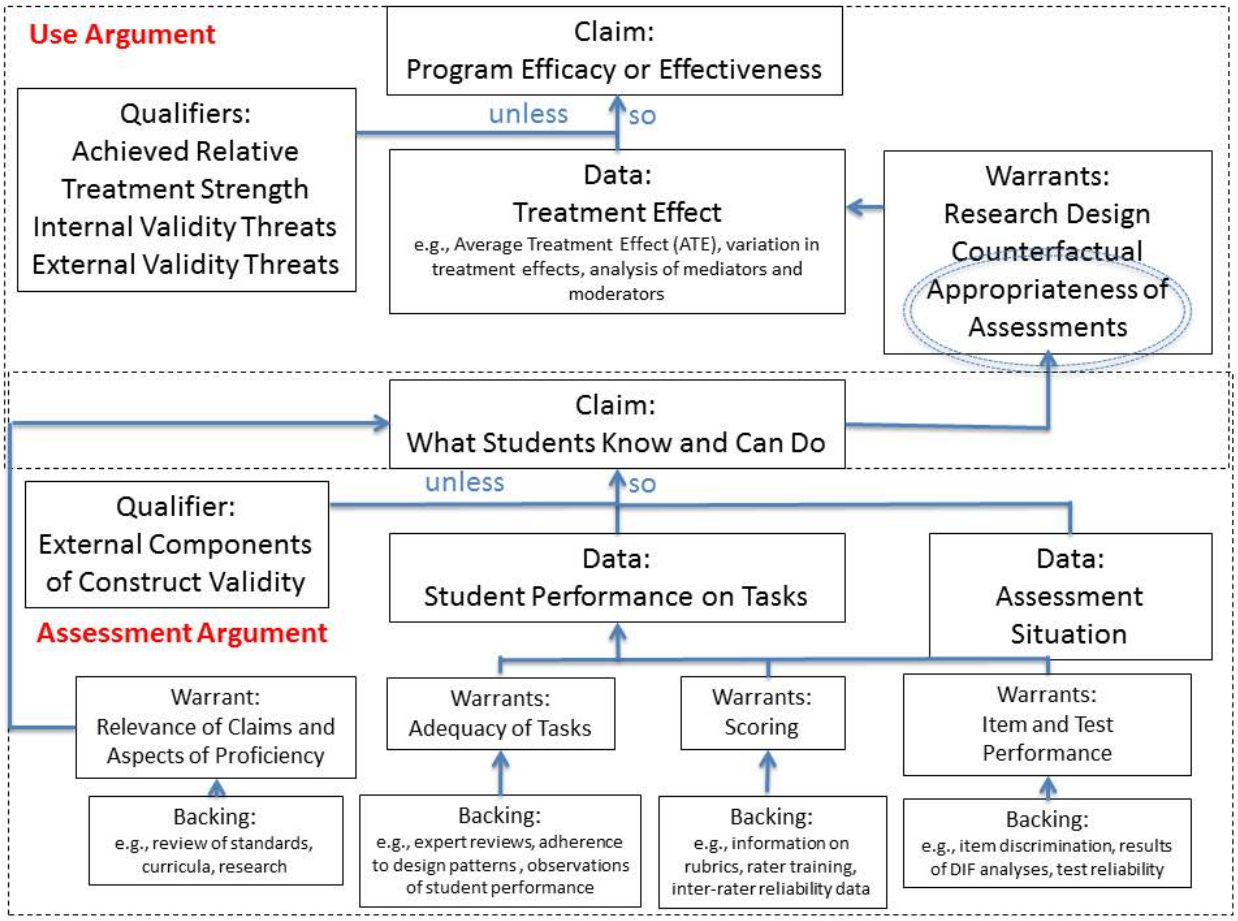


Figure 1. General structure of an assessment argument for evaluation use.

Even if data from student performance, warrants, and backing for warrants are compelling, assessors must still construct an argument for the use of assessment data for particular purposes (Messick, 1994; Mislevy, 2008; Shepard, 1993). In evaluation, a typical claim is an answer to a question about the efficacy or effectiveness of a program. Evidence that would support that claim comes not from one student, but from large samples of students in classrooms and schools. Whether or not those data support a claim about program effectiveness depends in part on the strength of the argument about the outcome measure to support claims about individual learning. But it also depends on the strength of the research design to support causal inferences (Shadish, Cook, & Campbell, 2002); the proximity of the content and tasks of the assessment to content and tasks presented to students the program (Ruiz-Primo et al., 2002); the nature and persuasiveness of the counterfactual to the potential users of evaluation results (Morgan & Winship, 2007); the quality of implementation, or what Cordray and Pion (2006) call the *achieved relative strength* of the intervention; and evidence regarding threats to

internal and external validity (Shadish et al., 2002). Finally, a compelling argument for the use of assessment for evaluation purposes considers possible counterarguments that would lead to different claims about effectiveness. For example, perhaps students performed better on assessments, but they did so because they spent more time on task content than the comparison group.

This broader structure of interlocking arguments underlies the design of the assessments we describe in this paper. Although the validity evidence we present focuses principally on supports for claims about students (the assessment argument), our design choices reflected the intended use of the assessment data from the start. We considered, for example, not only what claims about students we wanted to investigate, but also how to make the assessment a test for judging the efficacy of the curriculum that was sensitive to instruction but not "overaligned" to the content and tasks students encountered so as to reduce the credibility of our judgments. We have detailed the broad structure for our argument here, too, because we want to underscore the nature and breadth of evidence relating to validity that is needed to support uses of assessment for evaluation. The process of developing the assessment tasks we present in this paper was one that underscores the complexity and challenges that lie ahead for developing assessments of the performance expectations of the NGSS, whether these assessments are intended for use in evaluation or for other purposes.

From Performance Expectations to an Assessment Argument: Unpacking the Complexities

The NGSS performance expectations provide an excellent starting point for assessment design, yet moving from these to an actual assessment that will be valid for the purposes of evaluating a program requires building a coherent assessment argument that elaborates the claims, evidence, and reasoning to inferences about student learning that are desired. The following performance expectation serves as an example:

> MS-PS-1-4. Develop a model that predicts and describes changes in particle motion, temperature, and state of a pure substance when thermal energy is added or removed. *[Clarification Statement: Emphasis is on qualitative molecular-level models of solids, liquids, and gases to show that adding or removing thermal energy increases or decreases kinetic energy of the particles until a change of state occurs. Examples of models could include drawing and diagrams. Examples of particles could include molecules or inert atoms. Examples of pure substances could include water, carbon dioxide, and helium.]*

This performance expectation provides a lot of detail, such as how the science practice of modeling can be integrated with the core idea Structure and Function of Matter. Cause and effect is central, as students are required to use their models to think about how thermal energy affects particle motion. The clarification statement provides details about the kinds of models that can be expected and substances that may provide the context for items.

When we think about the claims we want to be able to make about what students know and can do, additional questions come to mind. For example, is there something qualitatively different about students' understanding and use of models when they are using them to predict or describe a phenomenon? If students are able to use a model (e.g., a provided representation) to describe changes in particle motion, would one consider this a component skill associated with this performance expectation or with a different performance expectation?

There also are pieces of the assessment argument related to evidence that need to be specified. What are the qualities that differentiate a strong model from a weaker one? What kinds of changes in particle motion need to be described—in drawing, in writing? In a task, do students need to provide evidence in the form of models for solids, liquids, and gases (or only for the effects of temperature changes associated with two of the three phases)? What is the level of detail with which atoms or molecules need to be represented in students' models?

Several questions also may be raised with respect to how student work will be scored and the nature of inferences made about student proficiency. Will there be multiple rubrics or a single rubric to capture the dimensions of core idea, modeling, and crosscutting concept for items associated with this performance expectation? Can scores on items associated with this performance expectation and items associated other performance expectations that relate to modeling be used to inform inferences about students' proficiency related to the practice of modeling?

In short, to use a performance expectation for assessment design requires further specification. The NGSS offer good guidance about the science domain and broad expectations; however, there are a number of design decisions remaining that must be formed into a logical argument to promote consistency in the design of the tasks within an assessment, and particularly so for assessments intended to inform program evaluation. Table 1 describes these decisions in relation to the warrants needed in the assessment argument (Figure 1).

Table 1. Questions to Guide Assessment Design Decisions

Warrants	Design decisions
Relevance of claims and aspects of proficiency	What opportunities are there in the curriculum intervention and in the business-as-usual curriculum to learn these core ideas, practices, and crosscutting concepts? What do we know about how children learn and develop? What state science standards are teachers accountable to? Given responses to the questions above, which Performance Expectations (PEs) should be targeted in the assessment? What are performance levels associated with the targeted core ideas, practices, and/or crosscutting concepts?
Adequacy of tasks	What are the design principles that guide the elicitation of intended claims/PEs?[a] Do observations of student performances provide evidence of the claims? If not, how should tasks be revised? Do experts agree that tasks are well aligned to claims?
Scoring	Are rubrics and scoring guides capturing the range of evidence related to the claims? Are raters able to consistently and reliability score tasks?
Item and test performance	What statistical approaches will be used to make inferences about claims from scores? Is the assessment a reliable measure of the claim(s)?

[a] In this project, we focused on core ideas, scientific practices and crosscutting concepts in the *Framework for K-12 Science Education* (NRC, 2012), because the NGSS were not available at the time we needed to develop our assessments.

An Evidence-Centered Design (ECD) Approach to Inform Decisions

ECD is one approach that facilitates addressing these design decisions. ECD requires up-front specification of student and measurement models to guide and promote coherence in the design of tasks and rubrics and the interpretation of student performances.

Two features are important to highlight with respect to the process of carrying out ECD. The first is that ECD requires codesign. To develop science assessments, ECD involves science educators, science content specialists, assessment designers, and psychometricians. Articulating the claims about what students can know and should be able to do must be guided by science educators and content specialists. Developing appropriate tasks and scoring guidelines requires not only domain knowledge but also experience with working with students for whom tasks are intended so that context and language are appropriate. Identifying which measurement models should be used to determine how student scores provide evidence of claims requires knowledge of statistics and psychometrics. For technology-

based assessments, software engineers also may be involved in the earlier domain modeling and conceptual assessment phases so that design specification are usable at the implementation phase. Importantly, codesigners coordinate their activity, but all interactions among designers do not take place within the full group.

Importantly, ECD is also an iterative process. Responding to these design decisions sometimes requires revisiting earlier decisions that have been made about claims and approaches guiding task design. If these models for assessment design are to be reusable and sustainable, they must be refined as new information is learned either from research about how learning occurs or from the performance of particular items and tasks.

Relevance of Claims and Aspects of Proficiency

In the context of the efficacy study of PBIS, one of the central research questions is: To what extent do students in PBIS perform better than non-PBIS students on measures of learning? We needed to develop measures that had potential to be instructionally sensitive in both conditions; thus we wanted to align our claims to standards—in this case, the core ideas, practices, and crosscutting concepts in the Framework. To ensure that we selected core ideas, practices, and crosscutting concepts that were fair to both conditions, we had to consider curriculum learning goals, state standards, and research on how students learn. In ECD, this phase of design is called domain analysis.

Curriculum learning goals. The process of identifying curriculum learning goals involved reviewing the curriculum units. In some cases, learning goals were explicit statements about content knowledge (e.g., Energy exists in different forms and can be changed from one form to another.) We also needed to review activities to determine whether students had opportunities to engage in particular scientific practices. Based on the curriculum analysis, the team selected several physical science core ideas, an Earth science core idea, and the scientific practice of modeling as areas of focus for the assessments. Selecting core ideas that are taught in the curriculum reflects a basic principle of fairness: students cannot be expected to know what they have not had opportunity to learn. As we elaborate later in the paper, the learning goals of the curriculum also figure in the evaluation use argument, and in a way that requires more complex thinking about alignment (or overalignment) to the treatment curriculum, as well as to the theoretical coherence of different treatments.

State standards. We learned that we also needed to consider state standards in defining and defending our claims. A challenge to our own study was that the NGSS had not been released yet or adopted by any states. While the district had great interest in the *Framework for K-12 Science Education* (NRC, 2012) and forthcoming (at the time) NGSS, they and their science teachers were accountable to the state science standards. Thus, alignment to the current state standards was an important consideration in assessment design.

Research on how students learn. As we settled in on particular core ideas and the science practice of developing and using models, we turned to the research literature to refine our conceptions

of this practice. For example, emerging learning progressions research (Schwarz et al., 2009) on modeling offered insights into the kinds of tasks teachers can present to students that can support students' engagement in modeling, in a manner that was consistent with the Framework. This literature stresses the benefits of having students construct and manipulate their own models, as opposed to working with preprepared models. It operationalizes the practice of modeling to include the following elements: (a) constructing models; (b) using models to make predictions or explain processes or phenomena; (c) comparing, critiquing, and evaluating models; and (d) revising models to better account for evidence. Moreover, this literature focuses on the progression in terms of students' abilities to engage with models in these various ways. Specific research on the practice within our targeted disciplines was also reviewed (e.g., Rivet & Kasten's [2012] development of a construct-based assessment in Earth science).

Defining claims and performance levels. Our work to define claims and performance levels focused on the practice of developing and using models, as we intended for this practice to be central in both the physical and Earth science assessments. On the basis of the research literature and the Framework, we defined four claims (or focal knowledge, skills, and abilities): (a) ability to construct a model and use the model to explain a phenomenon; (b) ability to construct a model and use the model to make a prediction about a phenomenon; (c) ability to evaluate the quality of the model for explaining a phenomenon; and (d) ability to use a given model to make a prediction about a phenomenon.

Two successive field trials informed the development of a construct map to describe performance levels associated with the practice of modeling (Table 2). The construct map serves to promote coherence in the way levels of proficiency related to modeling can be described in both content domains. Given that the target of our assessments is sixth grade students, the levels on the construct map span skills that at the lowest levels would be the focus of upper elementary school and at the upper level are expected in middle school (NGSS Lead States, 2013, Appendix F).

Adequacy of Tasks

To design and develop tasks aligned to our claims with respect to developing and using models, we began with a design pattern (Mislevy & Haertel, 2006). This is the domain modeling phase of ECD. We used the design pattern to develop tasks and rubrics and had experts in the field of science education review these. Expert reviews and an examination of student responses informed revisions to tasks.

Designing tasks to elicit the intended claims. The design pattern (Table 3) describes the argument underlying the design of our assessments—how we intended to elicit students' ability to engage in the practice of modeling using physical science and Earth science core ideas. The attributes in design patterns specify features of kinds of observations that can provide evidence about acquisition of a knowledge or skill, and the characteristic and variable features of task situations that allow students to provide this evidence (Mislevy et al., 2003).

Table 2. Developing and Using Models Construct Map

Levels	Level descriptors
4	The student recognizes models as a representation that can explain why a phenomenon is observed or that can be used to make predictions about the phenomenon. Model captures all mechanistic features of the observable and unobservable phenomena.
3	The student recognizes models as a representation that can explain why a phenomenon is observed or that can be used to make predictions about the phenomenon. Model captures some mechanistic features of the observable and unobservable phenomena.
2	The student recognizes models as a representation that can explain why a phenomenon is observed or that can be used to make predictions about the phenomenon. Model attends primarily to macroscopic, observable, or surface features with emerging understanding of mechanistic features.
1	The student conceives of a model as an analogy or an explicit representation of phenomena that is visible. Student-constructed model or student evaluation of a given model attends only to relationships among macroscopic, observable, or surface features to explain a phenomenon.
0	The student does not demonstrate any understanding of scientific models. Student-constructed model or student evaluation of a given model includes no appropriate relationships based on core ideas (mechanistic, surface, or otherwise).

The Framework does not provide specific guidance about, nor does it clearly differentiate between, latent knowledge and skills, student performances, and task features with respect to modeling. The design pattern schema was essential in this regard. Because the structure of a design pattern implicitly contains the structure of an argument in general, completing the design pattern simultaneously renders explicit the relationships in an assessment argument for developing and using models as a practice. In creating this design pattern, the team initially focused on the integration of content and practice. By investing in defining an assessment argument around a scientific practice, we were well positioned to apply the approach to assessment of modeling in both content domains. Importantly, while the claims (focal knowledge, skills and abilities [Focal KSAs]) in the design pattern highlight developing and using models, it is evident in the description of the potential observations and characteristic task features that core ideas and modeling must be blended in tasks. Over successive refinements to the design pattern, we incorporated the third dimension from the Framework: crosscutting concepts. Variable Feature 6 now highlights how mechanisms to explain phenomena through models may address micro- or macro-level relationships.

Table 3. Design Pattern for Developing and Using Models

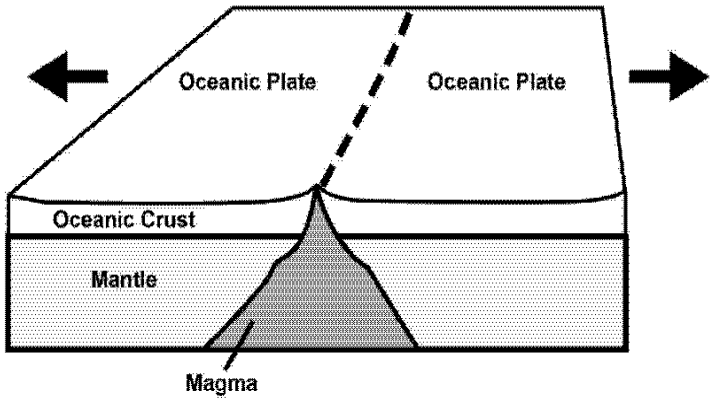
Attribute	Description
Focal KSAs *The primary claims targeted by the Design Pattern*	FKSA 1. Ability to construct a model and use the model to explain a phenomenon. FKSA 2. Ability to construct a model and use the model to make a prediction about a phenomenon. FKSA 3. Ability to evaluate the quality of the model for explaining a phenomenon. FKSA 4. Ability to use a given model to make a prediction about a phenomenon.
Additional KSAs *Other KSAs that may be required*	AKSA 1. Knowledge that a model explains or predicts. AKSA 2. Declarative knowledge related to core ideas. AKSA 3. Ability to construct a response in drawing or writing.
Potential observations *Qualities of student performances that constitute evidence of Focal KSAs*	PO1. Given a brief real-world scenario describing an observable phenomenon, student applies scientific concepts appropriately to construct a model (using drawings and words) that explains why the phenomenon occurs. (Physical science example: Given a representation of water molecules in solid form, student accurately constructs a representation of water molecules in liquid form and explains why water as a liquid can flow and change its shape to fit a container.) PO 2. Given a brief real-world scenario describing an observable phenomenon, student applies scientific concepts appropriately to construct a model (using drawings and words) and uses the model to make an accurate prediction about the phenomenon. (Physical science example: Student draws an accurate model to show how sound waves move from a ringing cell phone through the air and uses this model to make an appropriate prediction about whether students will hear the phone if it is put in a vacuum chamber.) PO 3. Given a model, student accurately describes similarities and differences between the model and a phenomenon. (Earth science example: Student identifies accurate similarities and differences between a cracked egg model and a scientist's model of Earth's surface/interior/geologic processes.) PO 4. Given a model, student uses the model to make a prediction about a phenomenon. (Earth science example: Given an image that shows the Hawaiian islands, the current location of the hot spot, and the direction of plate movement, the student correctly predicts the location of the next volcano and appropriately justifies his or her prediction using the model.)
Potential work products *Products produced by students*	WP 1. Drawing of a model. WP 2. Constructed response.
Potential rubrics *Potential scoring guides and approaches for evaluating student work*	PR 1. Rubrics must simultaneously distinguish among levels of sophistication with respect to both content knowledge and modeling.

Attribute	Description
Characteristic task features *Aspects of tasks that are necessary in some form to elicit desired evidence*	CF 1. All items must prompt students to make connections between observed phenomenon or evidence and reasoning underlying the observation/evidence. CF 2. All phenomena for which a model is developed must be observable (e.g., difference in temperature as a substance is heated, an erupting volcano) or fit available evidence. CF 3. Models provided in stimulus materials must illustrate a process or why a phenomenon exists (e.g., image of volcanoes over hot spot must include hot spot and direction of plate movement). CF 4. All items must elicit core ideas as defined in *Framework for K-12 Science Education* (NRC, 2012).
Variable task features *Aspects of tasks that can be varied in order to shift difficulty or focus*	VF 1. Drawing required: None, add to existing picture, construct model from scratch. VF 2. Complexity of scientific concept(s) to be modeled. VF 3. Format of "real-world" phenomenon presented: image, data, text, combination. VF 4. Core idea targeted in model: physical science core idea vs. Earth science core idea. VF 5. Function of the model: To explain a mechanism underlying a phenomenon; to predict future outcomes; to describe a phenomenon; to generate data to inform how the world works. VF 6. Scale of mechanistic relationships in model: Observable-macro, unobservable-micro, unobservable-macro.

Using observations of student performance to inform task design. The design pattern provided a common language and approach for item writers to implement modeling tasks and rubrics. In designing tasks, the team had to keep several constraints in mind. We had two sequential class periods (approximately 90 minutes total) for each assessment. Assessments needed to be delivered via paper and pencil, because it was not feasible to design and deliver technology-based assessments as part of this efficacy study. Using lab equipment also was not feasible because the project was not in a position to purchase and ship these materials to teachers. Thus, while we had a reasonable amount of testing time, we had the challenge of needing to design assessments that elicited modeling with more limited resources.

The task in Figure 2 illustrates several lessons learned with respect to designing tasks to measure the multiple dimensions of core ideas, science practices, and crosscutting concepts. The task intends to target Focal KSA1, “Ability to construct a model and use the model to explain a phenomenon.” In this task, students are asked to show what is happening inside Earth to explain the movement of plates and to use their drawings/model to support their explanations.

The picture below shows a place on the ocean floor where two plates are moving apart. At this plate boundary (shown at the dotted line), rock material is rising to the surface.



A. Draw on the picture to show what is happening in the mantle that causes the plates to move apart.

B. What is happening in the mantle that helps to explain why the two plates are moving apart?

C. Put an X on the places in the picture above where the oldest rock can be found in the crust.

D. Explain your answer.

Figure 2. Earth science task.

In earlier piloted versions of this task, students were asked only Parts C and D; however, we found that prompting students to explain did not elicit an understanding of the mechanism (model) about why the oldest rock was further away from the plate boundary. In the current version of the task, students still must indicate where the oldest rock is and use their models to support their explanations. We also revised the item based expert review, which revealed a key discrepancy between theories of mantle convection presented in curricula today and modern geoscientists' ideas about the role of the heating from Earth's core in driving convection currents.

Compared to earlier versions, the task is more explicit in asking student students to show/draw their models and in prompting students to indicate the components of a complete explanation using their conceptual models. We found that for middle school students, this level of scaffolding for eliciting their conceptual models was important so that students understood the kind of evidence we were expecting in their responses. Thus, many of the tasks on our Earth science and physical science

assessments have multiple parts. Figure 3 is an example task from the physical science assessment that similarly includes multiple parts to elicit aspects of model construction and model use.

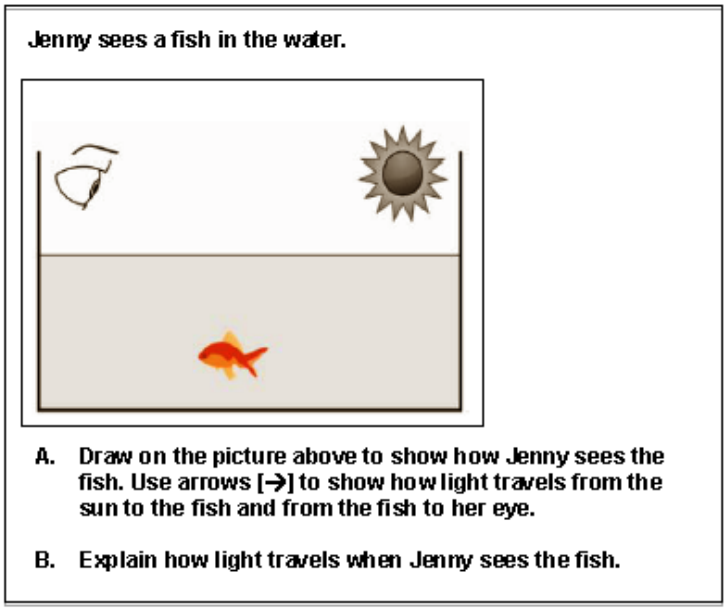


Figure 3. Physical Science Task

Expert review. As another check on the coherence between the intended design and the tasks, we asked experts in science education who served on committees to develop the Framework (NRC., 2012) and/or NGSS (NGSS Lead States, 2013) to review items and rubrics. Experts reviewed the 11 multiple-choice and 13 constructed-response items on the physical science test and the 16 multiple-choice and 11 constructed-response items on the Earth science test. For each item and its rubric, reviewers responded to the following questions: (a) Does the item assess a concept targeted by the core idea? (b) Does the item assess the practice of modeling? and (c) For ratings of "yes" or "partial" to the previous questions, how well does the item address the integration of content and practice? For any item judged as partial or not aligned, raters were asked to provide an explanation. Expert review has informed revisions to items.

Scoring

During scoring that took place for the piloting and field testing phases, we considered whether rubrics and scoring guides were capturing the range of evidence about the claims from students. We also examined the extent to which raters were able to consistently and reliability score tasks. These activities informed refinements to items and rubrics.

Evaluation of rubrics and scoring guides. Early scoring sessions with field test data involved a great deal of discussion about the quality of rubrics for evaluating student responses for evidence of the claims about modeling. These sessions were critical in this process of iterative refinement to both rubrics and items. We consulted an experienced psychometrician, along with middle school science teachers, to provide feedback on rubrics during these sessions. All scorers would review a sample of papers with the rubric as designed, check consistency in scores, and add detail to clarify or refine the rubric. After scoring all papers with the rubric, they also discussed ways that items may be revised in subsequent administrations to better elicit the desired evidence.

In the first field test, the assessment design team attempted to distinguish between content and science practice in rubrics to better understand the relationship between these constructs. Two rubrics were developed for each item: one that foregrounded content knowledge and the other that foregrounded developing and using models. While this exercise helped to tease apart elements of content and practice targeted by each item, the analyses revealed that content knowledge and modeling were difficult to disentangle and must be considered simultaneously in scoring (Kennedy, 2012a, 2012b). Blended content-modeling rubrics were better for interpreting student performances. Going forward, a single rubric (such as the example shown in Table 4) was designed to distinguish among levels of sophistication with respect to both content knowledge and modeling. For complete credit in the most complex items, students must construct drawings, explanations, and/or predictions that include scientifically accurate content knowledge *and* describe how their representation or drawing of the phenomenon helps to explain why a phenomenon exists.

In revising rubrics, we also worked to improve coherence between the rubrics and the modeling concept map. While each rubric blends content specific to the item with the practice of modeling, we also aimed for consistency across items with respect to the practice of modeling. Thus, we now have mapped each rubric to the construct map. Table 5 illustrates the mapping for the example discussed.

Table 4. Blended Core Idea—Practice Rubric for Earth Science Task

Score point	Descriptor for Parts A, B, C, and D
+1	A: Arrows are next to (or in) the magma on both sides angled up toward the crust, and then away from the magma and (maybe) down toward the bottom of the picture (the convection cycle). Arrows must be drawn in the mantle. All arrows drawn only up or down are not acceptable.
+1	B: Student explains that convection in the mantle drags/moves/pulls the two plates apart. (Student must talk about convection causing the plates to move, not just that the magma is "coming" or "pushing" up.)
+1	C: Xs are on both sides of the drawing, on the outermost edge of the crust. X's can be placed anywhere along outermost edge of crust, including multiple X's lining the outermost edge of crust.
+1	D: Older rock is dragged (moves) away from where the magma pushes up (plates move apart from the boundary). Or magma coming up between the plates and filling the gap with new rock.

Table 5. Mapping of Blended Rubric and Developing and Using Models Construct Map

Total score *See Table 4*	Level on modeling construct map *See Table 2*	Rationale
4	4	Student is able to construct an accurate model using drawing and writing that shows convection currents as the mechanism to explain the phenomenon of plate movements at a divergent zone. The explanation also includes an understanding about why older rock can be found further away from the plate boundary.
3	3	Student's model is mostly complete based on drawings and writings. Some aspects of the mechanism of convection currents are present. Details may be missing regarding aspects of convection currents or reasoning about where older rock can be found.
2	2	Student's model in drawing and writing is partial, with minimal evidence of the understanding of the mechanism of convection.
1	1	Student's model in writing and drawing attends primarily to aspects of the phenomenon of plate movement and or the location of older rocks. Evidence for the understanding of the mechanism of convection is not present.
0	0	No evidence of modeling.

Interrater reliability. A factor related to item quality is the degree to which different raters interpret student responses similarly. Scoring sessions typically involved four to five raters, and approximately 10% of all items were scored by the same raters. On both the physical and Earth science assessments, intraclass correlation coefficient (ICC) analyses were conducted. A two-way mixed effects model was used, with respondent values varying randomly and rater values fixed. These analyses revealed that interrater reliability was at least .80 on all items except for two Earth science items (Kennedy, 2012a, 2012b). This analysis signaled that for these items, revisions may be needed in the rubric (i.e., clarification of scoring levels). It also suggests that revisions may be needed to the item prompts; students may need better guidance about what is being asked.

Item and Test Performance

Analyses of item and test performance included item fit, item difficulty, and item discrimination were conducted to evaluate item and test performance. Fit of items on each assessment to a unidimensional Rasch model were examined to reveal gaps in representation of the construct. In this case, we were examining modeling in either physical or Earth science. Technical reports (Kennedy, 2012a, 2012b) provide details regarding item and test performance for the physical and Earth science assessments.

Assembling the Use Argument: Conceptualizing and Analyzing Treatment Strength

Our paper's primary focus is on the assessment argument, but in this section, we highlight design decisions with respect to select aspects of the evaluation use argument. We focus on aspects that are particularly relevant to the assessment argument and to the challenges unique to evaluation of curriculum materials in NGSS, namely the analysis of *achieved relative treatment strength.*

The concept of treatment strength is foundational in program evaluation. Treatment strength refers to the theoretical coherence of the treatment and how much of a treatment participants are expected to experience (Cordray & Pion, 2006; Sechrest & Yeaton, 1979; Yeaton & Sechrest, 1981). Analyzing treatment strength is important, because some treatments are likely to be so weak as to have little chance for success (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Relative treatment strength refers to the difference between the strength of a proposed treatment and that of a comparison treatment (which could be typical practice); analyzing it depends on analyzing components that are unique and essential to the treatment that is the focus of evaluation, as well as non-unique components that are essential (Cordray & Pion, 2006).

Program implementers change programs when they implement them in real settings. Sometimes, a treatment may not produce a significant effect because the integrity of the program has been compromised by those changes (Sechrest et al., 1979). Hence, it is critical for evaluation researchers to develop evidence about the *achieved* relative treatment strength, or the strength as

realized in implementation. In an evaluation use argument, achieved relative treatment strength is an important qualifier to inferences about the efficacy of programs.

In conceptualizing relative treatment strength for the PBIS evaluation study, our team faced a number of dilemmas. Early in the study, we did not know where we would be conducting our study. It was thus impossible to know what curriculum materials teachers in the comparison condition would be using. Without knowing what curriculum materials were in play, we could not compare the theoretical coherence of PBIS with comparison materials. Even after we selected a district for the study, we discovered through survey research that most all teachers regularly supplemented the district-adopted textbook with other materials. We did not have access to these materials to analyze their coherence, either.

Another dilemma particular to NGSS is that the developers of PBIS did not create curriculum materials to be aligned to either the Framework or NGSS. There are many opportunities to learn through engagement in the practice of explanation throughout, as well as opportunities for students to develop and use modeling. But several other practices are less prominent. In addition, the disciplinary core ideas that are focal in units chosen for the study (due to alignment to the state standards where we conducted the study) did not align fully to the disciplinary core ideas of the Framework. Our situation is not unique, and many evaluation researchers will face this dilemma in the future.

In response to these dilemmas, we constructed a curriculum theory of action that focused on what we hypothesized would differentiate teaching and learning in PBIS from teaching and learning in science textbooks that include few opportunities for direct investigation of phenomena. As part of this effort, we gathered input from curriculum designers about what they considered the key “active ingredients” of the curriculum. We also analyzed the district-adopted textbook, so that we could better understand how PBIS differed from it, in terms of how it presented content to students, including the opportunities provided for students to engage in science practices. Finally, we used the Framework as a guide to focusing on the kinds of opportunities to learn that we hypothesized would be linked to student learning. Here, we had to rely on the limited evidence base with respect to what teachers can do to engage students productively in science practices (e.g., McNeill & Krajcik, 2008).

Our constructs for implementation measures focused on a mix of unique and non-unique program elements that we hypothesized to be essential for improving teaching and learning. Data sources for these measures include weekly teacher logs, an annual teacher survey, observations, and teacher assignments and associated student work products. Each of these sources will provide us evidence related to the achieved relative treatment strength for PBIS. If those data indicate that achieved relative treatment strength is low, then any conclusions regarding the efficacy of the program must be qualified. Because our theory of action focused on comparisons to a generic, “typical” textbook, we cannot anticipate ahead of time how different the materials that comparison teachers use are from those used by PBIS teachers. The teacher assignments and associated student work data collections are essential for helping us to understand just how different those materials are.

Importantly, the measures we are using for analyzing achieved relative treatment strength are linked to the student learning assessments described in this paper. Specifically, our protocols focus on the same disciplinary core ideas and practices as do the assessments. They assess specifically the kinds of opportunities students have to engage in science practices of modeling and explanation, in the context of participation in activities focused on disciplinary core ideas in physical science and Earth science that are included on the assessment. The linkages will allow us to analyze evidence related to the instructional sensitivity of the assessments. We plan to examine the association between opportunity to learn and student outcomes among comparison groups, where teachers are using a variety of instructional materials, including the district-adopted textbook. Within the broader experiment, we can also test whether and how much teachers' engagement of students in practices to teach core ideas mediates any treatment effects. This analysis has the potential, too, to contribute to the field's understanding of how engagement in practices relates to student learning.

Discussion

Our assessment design process entailed many different considerations with respect to content and practices, context for the study, and the program we were evaluating. Our validity argument considered claims, data, warrants, and qualifiers for an assessment argument and importantly relates these to a use argument to promote valid claims about program efficacy or effectiveness. In the context of the NGSS, these interlocking arguments became more complex as not only core ideas, but also scientific practice and crosscutting concepts needed to be considered.

Our assessment design decisions were guided by ECD, which provided structure for defining and refining claims of performance expectations to develop tasks that blend core ideas, practices, and crosscutting concepts. ECD approaches were very helpful in building consensus among the design team about what it means to assess the scientific practice of modeling. In addition to the design pattern, a construct map served to promote coherence in the elicitation and scoring of modeling tasks.

We iterated on designs, developing and using evidence from scoring sessions and item modeling to refine construct maps, tasks, and rubrics. Cycles of early piloting and field testing were critical in providing evidence about the extent to which items were eliciting the intended evidence and rubrics were adequately capturing the critical range of ideas related to the claims reflected in the student responses. Importantly, these iterative refinements required convening teams with expertise in assessment design, science content, psychometrics, and science teaching. Our design process yielded a set of items that proved reliable and for which we have some preliminary evidence of instructional sensitivity.

One of the limitations of the assessments is that we focused on only particular components of the practice of developing and using models. Constrained by administration conditions, our assessments focused on dimensions of developing and using models. We were not able to easily elicit other components, such as model testing and model revision. Assessing dimensions of this science practice

may be better supported in formative classroom administrations, where students have more time to develop and revise their models, and with technology-supported assessments, and where conducting investigations and making revisions on the basis of data are possible. Nonetheless, this study demonstrates that, even under limited administration conditions, we were able to elicit aspects of this complex science practice.

At this time, the evaluation study of the PBIS curricular intervention is ongoing, and we are still developing evidence related to the use argument. The study is examining weekly online classroom logs as evidence of teachers' implementation of two curricular units, to understand teachers' enactment and the frequency with which they engaged students in science practices. Analyses of classroom video will provide evidence of how teachers' orchestration of class discussions, specifically talk moves, shapes opportunities for students to participate in the science practice of modeling.

Implications

Our aim in this paper was to begin to articulate a model for designing NGSS assessment for program evaluation. As we begin to see future assessment development guided by the NGSS, we need to not only ask questions about student performance data and evidence to support adequacy of tasks, scoring, and test performance. It is imperative to anticipate and design for the intended purposes for the assessment from the start.

While many of the assessment design decisions were specific to NGSS and the PBIS curriculum, the paper provides an approach for linking assessment and use arguments that may be applied in other evaluations of curricular interventions that intend to impact student learning. With the anticipated uptake of NGSS, new curricula that claim to promote student understanding of NGSS may be available. Studies are needed to evaluate the extent to which these claims are valid, and new assessments will be subsequently be needed, as existing measures are not aligned to NGSS.

In our work, tools like design patterns and approaches like ECD were critical in building consensus among the design team about how to design tasks to elicit evidence of claims. These kind of ECD-based schemas also may serve to build coherence in assessment systems. For example, a tool like the modeling design pattern that we developed is agnostic to purpose, and thus may be adapted to inform the design of modeling assessments for formative purposes. Thus, the up-front investment of generating these kinds of boundary objects can have potential long-term benefits in establishing coherence within an assessment system.

Author Note

This material is based upon work supported by the National Science Foundation under Grant Number DRL-1020407. 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.

References

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a means to link systemic reform and applied psychology in mathematics education. *Educational Psychologist, 35*(3), 179–191.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kennedy, C. A. (2012a).*PBIS student assessment on earth systems concepts: Test and item analysis*. Unpublished report.

Kennedy, C. A. (2012b). *PBIS student assessment on energy concepts: Test and item analysis*. Unpublished report.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B. B., Gray, J. T., Holbrook, J.,…Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning-by-design into practice. *Journal of the Learning Sciences, 12*(4), 495–547.

Kolodner, J. L., Starr, M. L., Edelson, D.C., Hug, B., Kanter, D. E., Krajcik,…Zahm, B. (2008). Implementing what we know about learning in a middle-school curriculum for widespread dissemination: The Project- based Inquiry Science (PBIS) story. *Proceedings of the 2008 International Conference of the Learning Sciences* (Vol. 3, pp. 274–281), Utrecht, Netherlands: International Society of the Learning Sciences.

Love, A. (2004). Implementation evaluation. In J. S. Wholey, H. P. Hatry & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed., pp. 63–97). San Francisco, CA: Jossey-Bass.

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching, 45*(1), 53–78.

Means, B., & Penuel, W. R. (2005). Research to support scaling up technology-based educational innovations. In C. Dede, J. P. Honan & L. C. Peters (Eds.), *Scaling up success: Lessons from technology-based educational improvement* (pp. 176–197). San Francisco, CA: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk, 2*, 237–258.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463–469.

Mislevy, R. J. (2008). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. In P. A. Moss, D. Pullia, E. H. Haertel, J. P. Gee & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 259–294). New York, NY: Cambridge University Press.

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

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2003). Improving educational assessment. In B. Means & G. D. Haertel (Eds.), *Evaluating educational technology: Effective research designs for improving learning.* (pp. 149–180). New York, NY: Teachers College Press.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference*. London, UK: Cambridge University Press.

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

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

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

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), 921-958.

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students ' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching*, *49*(6), 713–743. doi:10.1002/tea.21029

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

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., … Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654.

Sechrest, L. B., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. H. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. B. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Beverly Hills, CA: Sage.

Sechrest, L. B., & Yeaton, W. H. (1979). *Strength and integrity of treatments in evaluation studies*. Washington, DC: National Criminal Justice Reference Service.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Singer, J., Marx, R. W., Krajcik, J., & Clay-Chambers, J. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist, 35*(3), 165–178.

Toulmin, S. (1958). *The uses of argument*. New York: Cambridge University Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*(6), 716–730.

Yeaton, W. H., & Sechrest, L. B. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*(2), 156–167.

he Center for K–12 Assessment & Performance Management at ETS creates timely events where conversations regarding new assessment challenges can take place, and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state and local decision makers.

K-12 Center at ETS

Invitational Research Symposium on
Science Assessment