A Phenomenon-based Assessment System for Three-dimensional Science Standards: Why Do We Need It and What Can It Look Like in Practice?

William R. Penuel[1, 2]

Kim Frumin[3]

Katie Van Horne[1, 2]

Jennifer K. Jacobs[2]

[1]*School of Education*

[2]*Institute of Cognitive Science*

University of Colorado Boulder

[3]*Harvard University*

DRAFT: April 2018

Paper to be presented at the Annual Meeting of the American Educational Research Association, New York, NY

Corresponding author: Bill Penuel, william.penuel@colorado.edu

This material is based upon work supported by the National Science Foundation under Grant Number DRL-1748757, by the Gordon and Betty Moore Foundation, and by the Spencer Foundation (Award 201800036). 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 funders.

A Phenomenon-based Assessment System for Three-dimensional Science Standards: Why Do We Need It and What Can It Look Like in Practice?

The *Framework for K-12 Science Education* (National Research Council, 2012) presented a multidimensional design blueprint for new science standards, organized around a "three dimensional" vision for student proficiency. No longer were content and process standards to be defined separately as they had been in the National Science Education Standards (National Research Council, 1996, 2000) and in standards of other disciplines (e.g., Common Core State Standards in Mathematics). Instead, each standard was to be articulated as a "performance expectation" that fused together pieces of a disciplinary core idea, a science and engineering practice, and a crosscutting concept in science. As an example, a high school life science standard within the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) reads, "Apply concepts of statistics and probability to explain the variation and distribution of expressed traits in a population" (HS-LS-3-3). In this standard, the science practice is represented in the verb phrase, "Apply concepts of statistics and probability to explain" and refers to the practice of using mathematics and computational thinking. The core idea pertains to heredity, inheritance, and variation in traits, while the crosscutting concept of this standard is scale, proportion, and quantity and requires the application of algebraic thinking to analyze data and make predictions about the effects of changes on traits in a population.

Assessing these three-dimensional standards presents significant challenges to the field. For one, prior to the Next Generation Science Standards, most standardized tests measured how well students understood core ideas, without reference to either the science and engineering practices or the crosscutting concepts of science (Pellegrino,

2013). Second, it is not possible to assess even a single standard through a stand-alone item adequately; instead, assessing three-dimensional standards requires the use of multi-component tasks, that is, clusters of items, each of which is designed to elicit a component of one or more of the dimensions (National Research Council, 2014). Each item cluster or multi-component task, furthermore, should be organized around a natural phenomenon to be explained or a problem to be solved, as a means to support students' integration of the three dimensions as they respond to specific prompts (Achieve, 2018). Third, any summative assessments developed must be just one part of a balanced and equitable system of assessments that supports decision making at the classroom, school, and district level that provides coherent guidance to teachers about how best to prepare all students to meet targeted performance expectations (National Research Council, 2014; Shepard, Penuel, & Davidson, 2017; Shepard, Penuel, & Pellegrino, 2018).

Systems of science assessment that provide coherent guidance to teachers and that attend to equity do not yet exist, but there are efforts underway within and across states (e.g., Penuel et al., in press) and in school districts to develop and study them. This paper describes one such effort to build a system of science assessment from the classroom outward to the school and district level, as recommended in the National Research Council (2014) report, *Developing Assessments of the Next Generation Science Standards.* This emerging assessment system is unusual, in that it is centered on support classroom learning rather than focused only on alignment to standards as most assessment systems are. As such, it provides an appropriate context to explore how a system designed to be horizontally coherent can support coherent and equitable improvements to teaching and learning.

We describe this system as "phenomenon-based," because like the curriculum that is intended to support, phenomena play a key role in the NGSS. Phenomena here refer to "observable events that occur in the universe and that we can use our science knowledge to explain or predict" (Achieve, 2017). Phenomena—and their analogue in engineering, design problems—support integration of the three dimensions by providing a focal point for students' sensemaking. Centering phenomena and design challenges in curriculum, instruction, and assessment helps keep the focus on science as an enterprise organized around asking questions about the natural world and seeking to build theories and models to develop answers to those questions and engineering as beginning with problems, needs, or desires of human beings that need to be addressed.

We are in the early stages of studying this new, "phenomenon-based" assessment system, and it is still under development and operational only within a single grade level within the district. In this paper, we present a description of the system here, and we draw on interview and survey data from ten high school teachers in the district who are implementing the curriculum materials and making use of the assessments created to support implementation, to develop some preliminary conjectures about how and where our system is supporting teacher change and where it needs further refinement or significant rethinking.

The Importance of Coherence and Equity in a Balanced Assessment System

In a coherent and balanced assessment system, *coherence* has multiple, interrelated meanings. First, it means that all of the key actors in the system share a common vision of what improvement looks like, and the policies and assessments developed at each level of the system are organized around that common vision. When that is true, the system is

said to be "vertically coherent," because at whatever level we look in the system, we see people espousing similar ideas about how to improve teaching and learning (National Research Council, 2001, 2006). Second, coherence means that the key components that shape what teachers do—standards, assessments, curriculum frameworks, and professional development—all aim toward that common vision. When this is the case, the system is said to be "horizontally coherent" (National Research Council, 2001, 2006). Third, a coherent system is one in which people are engaged in ongoing work to refine, build, and test the guiding vision together. Just because someone judges guidance in standards, assessments, and curriculum frameworks to be aligned does not mean the system is coherent; coherence is an emergent and dynamic outcome of people working together both to "make sense" and "give sense" to current practice and how it needs to change, in order to achieve a particular vision for practice (Honig & Hatch, 2004).

Coherence of assessments can contribute directly to the quality of implementation of new standards and associated curricula. A number of studies of teacher professional development related to the "first generation" subject matter standards in mathematics and science found that teachers' perceptions of coherence were related to their implementation of instructional practices and curriculum materials aligned to standards (Garet, Porter, Desimone, Birman, & Yoon, 2001; Penuel, Fishman, Yamaguchi, & Gallagher, 2007). Studies also have shown the importance to teachers of coherence between external accountability assessments and classroom assessments with respect to their goals for learning (LeMahieu & Reilly, 2004; Wilson & Draney, 2004). Studies of teacher responses to professional development related to the NGSS similarly point to the salience of teachers' judgments of the coherence of assessments with messages about

instructional shifts of the NGSS (Allen & Penuel, 2015). This literature points to a consistent finding: teachers are more likely to make shifts in their teaching that embody new standards when they perceive guidance from external assessments to be consistent with the standards and the messages they receive about how to implement new standards.

Equity in systems of assessments is a critical but elusive goal within balanced assessment system. At a most basic level, it means that the assessments used to make inferences about student proficiency are fair and unbiased against particular groups of students, particularly those from nondominant communities (Camilli, 2013; Kunnan, 2004). Beyond the tests themselves, the uses of assessment data must also be fair, expanding rather than limiting students' opportunities to learn (Confrey, 2008). A focus on equity demands, too, that a balanced assessment system includes multiple means for gathering evidence not only related to the performance of different groups but on opportunity to learn (Guiton & Oakes, 1995; Herman, Klein, & Abedi, 2000; Smithson, Porter, & Blank, 1995). Evidence shows there is a strong relationship between student opportunity to learn on external assessments of student learning (e.g., Boscardin et al., 2005; Wang, 1998), underscoring the need for such assessments within a balanced system of assessments. The specific guidance in *Developing Assessments of the Next Generation Science Standards* (National Research Council, 2014) emphasizes the need for opportunity to learn indicators in science:

> Indicators of the opportunity to learn make it possible to evaluate the effectiveness of science instructional programs and the equity of students' opportunity to learn science in the ways envisioned by the new framework. States should routinely collect information to monitor the quality of the classroom

> instruction in science, the extent to which students have the opportunity to learn science in the way called for in the *Framework*, and the extent to which schools have the resources needed to support learning (such as teacher qualification and subject area pedagogical knowledge, and time, space, and materials devoted to science instruction). (p. 7)

Yet another aspect of equity in assessment systems pertains to the goal of helping students develop what might be called "practice-linked identities" (Nasir & Cooks, 2009; Nasir & Hand, 2006) in science. By practice-linked, we mean that students have a sense of connection between who they are and the activities or practices of science. Assessing links of classroom activities to interest, experience, and identity are not typically part of assessment systems, and yet creating such links is an important equity strategy for implementing the vision of *A Framework for K-12 Science Education:*

> [I]nstruction that builds on prior interest and identity is likely to be as important as instruction that builds on knowledge alone. All students can profit from this approach, but the benefits are particularly salient for those who would feel disenfranchised or disconnected from science should instruction neglect their personal inclinations. (National Research Council, 2012, p. 287)

A central challenge, then, for equity-centered assessment systems is to construct ways to assess the success of instruction in accomplishing the aim of building on prior interest and identity (National Research Council, 2014).

Building a New Assessment System for Science in Denver Public Schools

Since 2014, we have been working with colleagues in the Denver Public Schools to develop an assessment system that embodies the vision of the *Framework* for equitable

science teaching and learning. This is one of several lines of work within a long-term research-practice partnership (Coburn, Penuel, & Geil, 2013) formed in 2007 to support implementation of student-centered teaching approaches within the district. For this particular line of work, key partners are researchers and educators from the University of Colorado Boulder, Denver Public Schools, and Northwestern University. The system described below has been developed and implemented for high school biology, though plans to spread the curriculum-linked assessment system across the district are already underway with support from a new grant from the Hewlett Foundation.

Curriculum Materials

The assessment system directly supports student learning as part of a yearlong problem-based biology curriculum focused on the high school life science performance expectations of the NGSS. The curriculum is organized into three units: ecosystems, evolution, and genetics. Each unit is anchored in two phenomena and culminates in an engineering design challenge. Phenomena and design play a key role in making students' work purposeful and meaningful: they provide an anchor that guides a coherent sequence of lessons to support students figuring out science ideas and crosscutting concepts using science and engineering practices (Penuel & Reiser, 2018). In the three units that comprise the curriculum, students develop explanatory models for how trees can mitigate the effects of climate change, for how large herbivores have changed the ecosystem of the Serengeti, for how bacteria populations are evolving to become resistant to antibiotics, for how a junco population has become bolder across generations, and for how Duchenne Muscular Dystrophy is inherited and causes degeneration of muscles. And students complete design challenges in which they choose a tree to plant in their

schoolyard that will maximize biodiversity, design an infographic for a health clinic about why their behavior matters for slowing the pace of increasing antibiotic resistance, and hold a World Café in which they debate the merits of using new gene editing technologies to search for cures to genetic diseases.

The units present opportunities for student learning that represent a significant departure from how teachers have taught and students have learned in the past. The anchoring phenomenon is not just a "hook" to get students interested that is then left behind; students spend 4-6 weeks developing an explanatory model of a phenomenon and are expected to develop generalized ideas from encounters with similar phenomena that require application of the same disciplinary core idea. Lessons build upon one another in ways that are driven by students' own investigations and questions, instead of being modular activities focused on specific topics. The lessons each provide opportunities for students to use what they are learning in that lesson to add to an explanatory model they are building as a class for the phenomenon at hand, or to add some new knowledge needed to solve a problem given a set of criteria and constraints the students and teacher have defined together. Throughout, there are opportunities for students to connect what they are learning to their own interests and experiences, to raise new questions and explore them individually and together as a class. Finally, the culminating design challenges presented to students offer opportunities for students to connect science learning to ongoing community concerns and endeavors, to "break open" the encapsulated classroom and enable students to see how doing science might matter to them, to their class, and to the community.

Interest Survey to Select Anchoring Phenomena

The first component in the assessment system is one that is used to select anchoring phenomena for units. For each of the units, a design team comprised of science teachers, the district science coordinators, and researchers from the University of Colorado Boulder and Northwestern University undertook a systematic process to analyze a "bundle" of performance expectations (see Krajcik, Codere, Dahsah, Bayer, & Mun, 2014, for an example of such a process), brainstorm and explore the viability of candidate phenomena for their potential to help students figure out core ideas within the bundle, and solicit input from students about candidate phenomena. The process evolved over the course of a three-year period, but beginning in the second year, soliciting input from students through an interest survey became a central feature of the process.

The interest survey presented current ninth grade students with candidate phenomena in the form of questions that future ninth graders might explore. We asked students to rate how interesting each of the questions would be to ninth graders like themselves, as well as to rate how personally relevant the questions seemed to them. Typically, we presented between 6-10 possible ideas to students, ideas for phenomena we had already determined could adequately address the bundle of performance expectations we had selected for the unit. Between 150-200 ninth graders across the district completed the survey each year, and results were disaggregated by gender, race, and home language to identify patterns that could help us identify phenomena that would be interesting and relevant to students from a wide variety of backgrounds.

We consider this interest survey to be a component of the assessment system, and not simply a part of our curriculum design process, because it relates directly to a core

assumption of *A Framework for K-12 Science Education* (National Research Council, 2012), namely the idea that science instruction should connect to students' interests and experiences. The interest survey provides a partial evidentiary warrant that the focal point for the science curriculum is interesting and personally relevant—at least to students who might be similar to those in design team teachers' classrooms in future years. It is hardly sufficient, though, to ensure that future students who encounter the materials will find the material interesting and relevant; additional components of the assessment system teachers attempt to address this need.

Routines and Artifacts for Eliciting, Prioritizing, and Answering Student Questions

The assessment system includes a number of teaching routines (DeBarger, Penuel, Harris, & Schank, 2010) and artifacts that can support teachers in making use of student questions to motivate classroom investigations of phenomena. These routines support multiple goals simultaneously: eliciting prior knowledge, interests, and experience that can serve as resources for instruction; creating a focus for teaching; developing students' grasp of the practice of asking questions in science; helping students keep track of progress they are making toward answering the Driving Question (Krajcik & Mamlok-Naaman, 2006) for the unit. The artifacts are public records of student questions, as well as places for them to record answers to those questions derived from investigations.

One such routine is the *Anchoring Phenomenon Routine* developed by our colleagues at Northwestern University (Reiser & Novak, 2017). In the anchoring phenomenon routine, the teacher presents the phenomenon for the unit to students in the form of a demonstration, an investigation, or a video. Students write down what they observe or notice, and they write down and discuss their questions as well. The teacher then creates

a public record of what individual students noticed and the questions they had. Next, students work in small groups or individually to generate initial explanations for what they saw. Note that they do so well before they could be expected to construct a complete explanation for the phenomenon; this step both provides the teacher with a sense of what knowledge students bring and helps students generate a sense of what they do not yet understand about the phenomenon at hand. The process of building a public record of areas of consensus and disagreement that follows this activity further facilitates question generation.

The next phase of this routine presents students with the opportunity to make use of their own experiences, focused on the question, "Where else does something like this happen?" The teacher asks students to think of experiences they have had that might be related to the phenomenon at hand and how they might support their making sense of that phenomenon. The class builds a public record of these experiences, adding questions students might have about these experiences that they might pursue—either individually or as a class—over the course of the unit. This step is critical to ensuring that the phenomenon connects with the students in that particular class; the related phenomena they generate will inevitably vary from period to period in a secondary classroom, and from year to year in all classrooms, introducing the need for the teacher to be prepared to support different learning pathways through the unit.

In the final phase of this routine, the class creates an artifact called a Driving Questions Board (DQB) that sets the initial direction for the unit as a whole. The DQB is a jointly constructed artifact that reflects the questions that the class believes will help them make sense of the anchoring phenomenon and answer the Driving Question

(Weizman, Shwartz, & Fortus, 2008). The class generates a Driving Questions Board by asking individual students (or pairs of students) to again frame a question, this time that the class can take up as a whole to make progress in answering the Driving Question or explaining the anchoring phenomenon. Students each have an opportunity to post their question to the board and defend their question. The teacher and student then group and prioritize the questions, focusing on an order that the class agrees is logical for pursuing an answer to the Driving Question. This collective effort is intended to build ownership over the Driving Question and the associated student questions (Weizman et al., 2008). As a final step in the routine, students outline some foci for investigations they might pursue in the class to answer the questions they have generated.

When the anchor is a design challenge, the routine follows a similar sequence, but is adapted slightly. For example, instead of a phenomenon, a problem is presented for students to solve. And in addition to generating questions, students also define a set of criteria and constraints for solving the problem they have been presented. Sometimes these criteria are partly given in the presentation of the problem, but other times the students must generate them themselves, and those criteria and constraints evolve and are tracked over the course of the unit.

Another routine developed by our partners at Northwestern is the *Navigation Routine* (Reiser & Novak, 2017). This routine is an essential tool for supporting students in connecting lessons in a unit. It begins with teachers asking students what they figured out last time about the phenomenon or problem, and then asking them to reflect on what questions they decided they needed to tackle next, to make progress in answering the unit Driving Question or one of the questions on the class DQB. Then, the teacher and

students partner to decide how to pursue one of those questions, guided by the teacher toward one of the investigation that curriculum developers have created for the unit.[1] At the conclusion of the day's investigation—which may entail analyzing data, obtaining information from scientific texts, or planning and conducting an explanation—the teacher brings the routine to a close by asking students to say what they figured out about something on the DQB and to choose a question to take up in the next day's lesson.

This particular routine supports students in individual and group self-assessment. Students in the past have said that returning to the DQB on a regular basis helps them connect ideas across lessons and focus the learning on questions they wanted to answer as a class (Weizman et al., 2008). In the units we have designed, this routine is intended to facilitate intra-unit coherence from the *student* perspective (Reiser, Novak, & McGill, 2017), that is, to help students understand exactly why they are doing that day and how it is building toward an explanatory model of the phenomenon or a solution to the problem the class has been presented. It also engages them fully in an ongoing way as partners in co-defining the aims and direction for the day's lesson.

Student Electronic Exit Tickets

On a weekly basis, students complete brief exit tickets that are embedded within the curriculum materials. These exit tickets are intended to function as a kind of "practical measure" (Yeager, Bryk, Muhich, Hausman, & Morales, 2013) of student learning and experience—that is, a measure that is feasible to implement within the flow of instruction and that can inform the teaching and learning progress. The intent is for teachers to be able to use these to assess the degree to which students were able to add to their

understanding of the phenomenon at hand what was intended in the written lesson, as well as to measure student experience of the curriculum.

Students report on different aspects of their experience in the exit tickets. They report on its perceived coherence, including their clarity about why they are engaged in the day's lesson and how it contributes to them answering the unit's overall driving question. In addition, they report on the perceived relevance of the day's lesson, indicating whether it mattered to them, to the class, and to the community. Third, students report on their affective response to the lesson, that is, whether they felt bored, excited, confused, or confident during the lesson. Our research to date indicates that the coherence of students' self-reported learning experiences are associated with feelings of excitement and perceptions of relevance to their lives and their community, two important dimensions of student engagement (Penuel, Van Horne, Severance, Quigley, & Sumner, 2016). We view the exit tickets as a key component of the assessment system, because it provides evidence as to students' disciplinary learning and their perceptions of how connected the lessons are to one another and to their own lives.

Routines and Artifacts for Building Explanatory Models of Anchoring Phenomena

As noted above in our description of the curriculum unit, the anchoring phenomenon is not merely a hook to introduce science ideas to students; instead, it is a thread that connects lessons to one another. But students need support in putting pieces together to build an explanatory model of a phenomenon—that is, a model that includes the key components, interactions, and mechanisms that account for what students have observed about the phenomenon. The assessment system therefore incorporates routines and

artifacts to support student sensemaking, one of which links directly to a set of routines and artifacts that the district has put in place to support teacher evaluation.

One curriculum-embedded routine is the *Putting Pieces Together Routine* (Reiser & Novak, 2017). In this routine, which takes place after several connected lessons in which students have developed fragments or pieces of models of the anchoring phenomenon, the teacher facilitates a discussion in which the class takes stock of what the progress they have made over the past several lessons. Students develop and share public representations of their models of the phenomenon, and they may use a "gallery walk" participant structure (Kolodner, Crismond, Gray, Holbrook, & Puntambekar, 1998) to compare and contrast models that different groups have generated. Then, in a discussion, the class comes to consensus on an explanatory model for the phenomenon. The teacher may then follow up the class consensus with an individual assessment, in which students are asked to construct a model of the phenomenon and support their model with evidence from the investigations the class has conducted.

To ensure that this process of coming to a consensus model is as efficient as possible, alongside the DQB, we have developed a "model tracker" artifact that students can contribute to on a regular basis. This artifact is intended to help students keep track of and organize the pieces of models they have developed into a format that scaffolds their sensemaking. Its integration into the assessment system addresses a well-known challenge with problem-based learning, namely students' ability to hold in working memory pieces of a complex model they are developing (Kirschner, Sweller, & Clark, 2006).

Teachers can evaluate their students' individual models using a set of rubrics we have developed that can be used to support the district's process for measuring progress toward the accomplishment of Student Learning Objectives (SLOs). SLOs are used in many states and districts as an alternative measure of student growth for use in teacher evaluation; defining an SLO typically involves the development of measurable targets for student achievement that follows an analysis of baseline data (Crouse, Gitomer, & Joyce, 2016). In Denver, teachers are responsible for negotiating the focus of their SLO with their principal, following a district-provided format for setting objectives and gathering evidence of student learning. Our partnership has developed rubrics that follow the district format and that can be adopted or adapted by teachers. Teachers can use rubric scores derived from analysis of student models of phenomena in units as evidence to support claims about growth in students' grasp of the practice of developing and using models.

Transfer Tasks

A key learning goal is to support students' generalizing from their investigations of phenomena. It is not enough that students be able to explain a particular phenomenon; it is critical that they be able to abstract science ideas that they can apply to the study of related phenomena. Typically multiple cases are necessary, in order to facilitate reasoning from cases to develop generalized ideas (Kolodner, 1993; Kolodner, Gray, & Fasse, 2003). For this reason, our units present more than one phenomenon to students, and students are supported in applying ideas from one to a subsequent phenomenon and elaborate on those ideas. In addition, we have developed a set of summative assessment tasks that teachers can use that present students with an unfamiliar phenomenon in which

they must use science and engineering practices to explain, applying focal core ideas and crosscutting concepts.

The design of the transfer tasks follow guidance regarding the construction of multi-component tasks organized around a scenario that presents a phenomenon or problem to students (Achieve, 2018; National Research Council, 2014). As an example, for the evolution unit, one transfer task presents evidence of change within a population of swallows that adapted to life beneath a new interstate highway. Students analyzed data about wing length, nests, and road kill to build an evolutionary explanation for how the population might have changed. The students apply what they have learned from two related phenomena involving bacteria and a different species of birds, to answer these questions.

These transfer tasks are integrated into optional unit tests that are provided to schools by the district central office, for the purposes of external monitoring. Through our partnership, we have been able to integrate these tasks into those assessments through the typical review process used to develop these assessments. The district assessments include other tasks, however, that have been developed by teachers not using the units and with more limited preparation in the design of three-dimensional assessments. To support more coherent assessment development, members of the partnership team have developed professional development supports for district teachers writing assessments, which we are currently studying.

The Current Study

Our primary research questions for this initial study of teachers' perceptions of this system in development were:

1. How do teachers perceive the assessment system components and how they fit together?
2. What personal experiences and aspects of their school and the district context inform their perceptions?

Sources of Data

The primary source of data for this analysis are two sets of interviews conducted with ten teachers from the project, half of whom had been involved in co-designing the curriculum materials. Of note is that even co-design teachers were less involved in designing the assessment components than were lessons, with the exception of the embedded assessments where students develop models of anchoring phenomena and with opportunities to review and revise exit tickets.

The interviews were conducted over the course of 2016-17 and 2017-18 school year by the second author, who is external to the partnership and serving as an ethnographer to the partnership. They are part of her dissertation study on learning in research-practice partnerships. All data that shared with the partnership was de-identified and redacted to ensure complete anonymity. Interview protocols in the first year focused primarily on what people learned from the partnership, and also on partnership challenges. Assessment was not a primary focus of the study, but assessment emerged as a concern and area of focus for the partnership during that year. So, for the second year, we therefore agreed to include on the interview protocol two questions related to the concerns teachers had raised with us in professional development and to our own research interests: What do you do for grading in iHub units? Do you develop your own assessments, or do you use the district ones?

A secondary source of analysis is an a still-in-progress survey of teachers' perceptions of the different systems components. We examined responses to six teachers who have completed the survey so far (19 were sent the survey), two of whom are co-design teachers. Because the survey is still in progress and does not represent adequately the full range of views of the field test teachers. relied on these responses to triangulate rather than as a central basis for interpretation. This survey included questions about the use of different assessments, as well as their perceptions of their value. It included also an open-ended question about what they most like about and would like changed about the assessments.

Approach to Analysis

Overall, we characterize our approach to the analyses of these interviews as are taking an *actor-oriented* perspective on teachers' perceptions of, orientations to, and implementation of components of the assessment system. In the context of this larger endeavor, in which we aim to make use of implementation evidence formatively for iterative design, this perspective helps us support more effectively teachers' learning needs, by building on teachers' resources for making sense of the assessments and thinking rather than look at it mainly from a deficit perspective. An actor-oriented analysis focuses on how teachers interpret guidance embedded in materials and how these perceptions shape their decisions about how to adapt materials to their local circumstances. Originally conceptualized as a model for understanding transfer in students' subject-matter learning (Lobato, 2003, 2006), an actor-oriented analysis of curriculum enactment seeks to produce an account that links teachers' decisions about implementation to what they interpret to be salient curricular purposes and structures and

how these interpretations are shaped by prior experience and their local context (Penuel, Phillips, & Harris, 2014).

We began by identifying excerpts where assessment mentioned in the first years of interviews, and then for the second year, we selected responses to questions that asked specifically about assessment and teachers' perceptions of the system components. Then, we engaged in coding to identify the specific components discussed, followed by a thematic analysis related to the codes. Finally, we engaged in a second layer of more inferential analysis of the resources for sensemaking teachers were bringing to making sense of the components.

Through this process we identified a total of 19 different excerpts from interviews related to assessment, though not all of the components of the system were referenced in interviews. We report on those here that were, along with components that teachers mentioned but that not formally part of our designed system.

Results

Embedded Culminating Assessments: Models of Phenomena and Design Proposals

The most commonly cited assessment component in the system were embedded models of assessment. These were characterized generally positively, with phrases like "remarkably rigorous" and "authentic." One of the tools that teachers use to make sense of assessments that integrate science practices is a format for written argument referred to as a "CER," which is short for claims-evidence-reasoning. The district has led professional development in the use of this tool to support the practices of argumentation and explanation, drawing on the work of our colleagues in this area (McNeill & Krajcik, 2012). In this particular example, the teacher also uses the idea of a "scientific paper" to

make sense of what we have introduced in professional development, perhaps relating to a different practice—obtaining, interpreting, and communicating scientific information:

> So that means that we're like creating a poster that we're presenting, like you would do at a conference or you're basically writing up some kind of CER is what we'd call it. Like some kind of claim-evidence reasoning. It's basically like a conclusion to a paper where, I mean, you know, it's kind of like expository writing I guess but it's like facts and argument kind of slam together. We're really trying to push the kids towards that because that's what you would publish in a paper.

Of note is that this particular teacher sees this as a worthy direction for them and their colleagues, but also something they are just beginning to explore.

One teacher reported on adaptations they had made to activities in which students constructed and evaluated one another's models of phenomena. This teacher was particularly pleased with the activity they had adapted, both because of what they observed in their students, and also because there was a district observer conducting a formal evaluation of them on using the district protocol:

> Yeah, like developing a model, that's what it is, and so students would develop models, but then after they develop a model, there is no piece to evaluating other models afterwards and seeing what other groups did, so we came up with something where we did a Google form, and students would evaluate anonymously other posters, talk about what went well, what they could have improved on, what you would give them out of a grade of 10, and students would actually fill that out on their phones, and then I would auto-populate a spreadsheet

> where they could immediately see how they did and what other groups are seeing. And so when I was being evaluated, I was actually doing that, and that's when I got like the highest score I've gotten, because of using that strategy so students—I was using technology in the classroom. The students could see what they were doing well and what they could improve on in an anonymous way. So that lesson went very well, and I do like the use of developing models too.

In the initial part of this response, the teacher struggles to recall the name of the science practice, and later admits they didn't know what a model was at the beginning of the year, but the assessment examples embedded in the unit that focused on building models helped them and their students grasp the practice better. In this example, there's strong congruence between the intentions of the developers of assessments with the assessment; it would be fair to say this adaptation enhanced the written activity, because of the anonymous peer feedback introduced.

Some teachers complained about the time it took to grade models, but they valued the evidence from models. One said that although it took "hours and hours" to go through and were "insanely time-consuming to grade," the assessment data were, they said, "way more relevant than just like a multiple-choice quiz." Here, the resource the teacher brings to make sense of what's valuable about the assessment is actually a contrast – to multiple-choice quizzes, a form of assessment which they are more familiar. Another teacher who said "that they take way too long to grade" offered the idea to the team that student models could be incrementally built and assessed every lesson:

> Each lesson, at the end of each lesson, they draw a model, and then they do lesson 2, and then they draw a model at the very end, and one thing that I think would be

> great is if we had the students draw their model for homework. It only takes five, ten minutes, and then they come in that next day, and then we have a rubric up on the board for each model and say, all right, out of ten points, how many did you get? And these are the ten things that I'm looking for in your picture, and you have to like trade your models around the room and check different things, maybe do a few different rotations, but I think that that is a great way to hold students accountable and to track how they're doing lesson by lesson, that I'm looking forward to teaching that next year.

In fact, we did create a "model tracking tool" that was intended to support incremental model building, and half the teachers who completed the survey were both using this tool and were finding it useful for both assessing student work and informing their ongoing teaching.

In addition, survey data confirmed the pattern in the interview excerpts observed – that these tools were most commonly mentioned by teachers as ones they use and find very useful for different assessment purposes, from engagement to analysis of student learning.

Transfer Tasks and District Assessments

The most commonly mentioned component after the embedded assessments were transfer tasks in which students were asked to explain a new phenomenon. Some teachers appear unaware of our involvement as a partnership in helping to develop district tests. Though these tests are not entirely comprised of multicomponent tasks we have written, they typically include two to four such tasks. One teacher said they hoped the project

would develop “item banks” of tasks to draw from, because they did not have much trust in the quality of the district tests:

> There’s a lot of pushback with the benchmarks with that where the questions are picked out and teachers, there’s only a selected few that volunteer to do it, but then the other 80% are complaining that they’re not good questions, whereas a test bank of questions and allowing teachers to choose those I think would be the best option.

Others—who may or may not have been aware of our partnership’s involvement in supporting changes to district assessments, believed the units prepared their students well for district assessments. They saw the link in terms of phenomena that anchor the assessments, just as they do the units:

> I think that they would do better on district assessments, because they can actually have an understanding of what the processes mean and relate it to different phenomena, and so that’s what district assessments are is different phenomena, same concept, so I think that they would do better.

Student Electronic Exit Tickets

Only one excerpt from interviews addressed our student electronic exit tickets. The one reference was a positive one, and the teacher referred to them as formative assessments: “We can use those as formative assessments. So they're kind of providing us with those tools and just trying to learn how to implement those formative assessments sort of in a smooth, transitional kind of way.” The limited focus on these exit tickets is consistent with survey data, which indicated less use of these tools and that they were for the most part only moderately useful as assessment tools.

Other Components Mentioned by Teachers

Teachers mentioned a number of assessment components that were not part of our designed assessment system. For example, somewhat unexpectedly, some teachers mentioned activity sheets that students complete during lessons as important formative assessment evidence for them. The implementing (i.e., not co-design) teacher below from the first-year interview commented these provide both information about whether students are struggling, and refers to the sheets as "experience-based," something that puzzled us as analysts but bears further investigation:

> There are a lot of constructive-response questions that I'm able to go around the room, and if they're not writing anything, I can ask them questions to phrase in a different way, and pretty much every student will have an answer to every question, so I think that's good. And last year it was kind of like cut and dry, this is the answer, next question, and this year it's much more experience-based, like what I'm seeing about my students is my students' experiences with this material.

Note the language of assessment (constructed response) here is mixed in with more informal language that attends to equity of participation ("every student will have an answer to every question") and to how students are interacting with and learning from the material.

Another aspect of assessment—not a component per se, but an embodiment of a principle championed by advocates of formative assessment, namely the clarification of learning goals for students (e.g., Black & Wiliam, 2009)—mentioned by teachers pertains

to the development of Content Learning Objectives (CLOs). CLOs are a district requirement for lesson, and they are part of what teachers are formally evaluated on (i.e., Do they post them on the board and discuss them with students?). They require a particular form that many teachers find cumbersome, as this one did, and contradictory to the idea of students "figuring out" the science that explains a phenomenon:

> One of the big struggles is like a content language objective. So we're supposed to have an objective at the beginning of class. This is what people need to know, but with the way this works, we can't tell them what they need to know until they figure it out, and so goes against what we're trying to do with the inquiry model when we're required to tell them what they need to know. So we either have to write really weird, vague-sounding content language objectives, or not do it and lose those points, or do it and ruin the inquiry part of it.

Note the language of "practices" (part of the *Framework*) is the language of "inquiry" but consistent with our partnership's intentions. To us, this excerpt suggests both the need to refine and support teachers in language for describing the approach but also provide them with tools for developing CLOs that meet district requirements without giving away the science. One teacher this year has created a process for developing CLOs collaboratively using the Driving Questions Board as a resource for doing so. Their students build CLOs around what they are figuring out, rather than around the disciplinary core ideas that students actually figure out in the course of a lesson.

Another teacher raised a similar issue about the mismatch between the demand for a CLO focused on "content" and the emphasis on figuring out content through science

practices. Here, the language of "practices" is a key resource for how they adapted the lessons to develop a CLO that met district requirements without giving away the science:

> The major tweak that I see that needs to be done, and I do for myself, which is big for a LEAP rubric, is connecting your CLO with what you end your day with, if it's a reflection or if it's whatever. So just making sure those connect and the students show some sort of growth in there, and, again, I'm not saying mastery by the end of the class, but it needs to be very specific, so you just have to write a really good CLO. And, again, normally my CLO is based on standards in the SEPs, so science and engineering practices, because I know they're doing that the whole time, so then what is the content? Well, my content is the science and engineering practice, and how they are going to do it. They're going to write in complete sentences, or they're going to verbally say evidence, or something along those lines.

Some teacher-made assessments were not productive adaptations, though they were responses to external pressure from parents to yield products that produce points and grades. For example, one provided vocabulary quizzes, things that were "relevant to what we're doing, but maybe not really covered in the lessons that we have from iHub. They explained, "I also do things that just give me some points to put in the gradebook, because there aren't enough assessments, and I've had my tail chewed by parents saying, you don't do enough assessments." Notably, the implication here is that the curriculum itself is insufficient, in terms of its coverage of the standards, for this teacher, and so they are creating assessments that are misaligned with the approach advocated by designers, as well as likely supplementing the instructional materials. This teacher also felt that if the

team got more input from teachers on the assessments in their design, that they would be more usable to them.

Discussion and Conclusion

In this paper, we have described an effort to bring about a new kind of coherent and equitable system of assessments of next generation science learning. These assessments, informed by sociocognitive and sociocultural theories of learning, are varied but linked to a common vision of teaching and learning outlined in a *Framework for K-12 Science Education* and embodied in the NGSS. The effort is not only a work in progress, but also—as our analyses of teacher responses to the system—in need of refinement and recalibration in some instances. Many of the reasons for needing recalibration arise from the practicalities of teaching large numbers of students, as well as from pressures outside the direct influence of our partnership (e.g., parent expectations, district mandates regarding CLOs).

Teacher agency is both evident in their responses to challenging circumstances, and there is room for us to allow for more to improve the assessments. On the one hand, teachers' adaptations have and will continue to inspire innovations to help teachers integrate these assessments more readily into their practice. The model tracking tool is one striking example, and the collaborative construction of CLOs with students is another. But more than one teacher expressed a desire to have more of a hand in their design. Our presentation of assessments—in contrast to curriculum materials—is of mostly complete assessments teachers are invited to give feedback to improve.

Strikingly, district assessments over which we have had some influence remain for most teachers problematic. It may be that it is the fact we only partly influence the

content of these assessments as a partnership that accounts for this finding. Many of the writers in the district have had less exposure to the NGSS and to our materials. Though all have received some basic professional development in writing three-dimensional science assessments, many have not experienced teaching or encountered materials that reflect the vision. Our curriculum units are a district initiative but remain for now "opt in" materials for schools, in keeping with the district's philosophy about autonomy.

With respect to the ways that teachers interpret these assessments, we are struck by the ways their interpretations are consistent with the vision of the Framework with respect to the centrality of phenomena and science practices. Missing, however, is talk of crosscutting concepts or themes of equity regarding the value of assessing the degree to which lessons matter to students and are exciting and engaging. While some teachers talked about how engaging the phenomena were, they did not value the student electronic exit tickets—which assess engagement—as much as the other assessments.

We are already engaged in studies to address some of the issues teachers have raised. For example, we are embarking on a series of studies focused on helping a broader group of teachers learn to design three-dimensional, multicomponent assessment tasks. The intent is to integrate these into district assessments and, where appropriate, the curriculum itself. And, we are embarked on new validity and usability studies of our exit tickets. One aspect of the study will be to foster teacher team discussion of engagement data, in order to explore how these data might be used formatively.

At the same time, we recognize that teacher adaptations in response to pressures from other system components over which our partnership has little to no influence will continue to be needed. Our efforts to create a completely coherent system for teachers are

limited by time, resources, human capacity, and influence. Some challenges arise from the decoupling of processes within the district itself as well—such as the separate processes for curriculum design and assessment design. There is, we believe our analyses show, a limit to what a partnership can accomplish in terms of coherence, and it is a reminder that coherence is a form of "craftwork" that requires actors at different levels of the system to coordinate activities, adapt, and learn from each other (Honig & Hatch, 2004).

How This Work Is Informing Others

One aim of a research-practice partnership like ours is to inform the work of others, both in research and in practice. Within the research community, the system we are building is serving as a kind of "worked example" of how assessment can work practically to support teachers (Heritage, 2018). Other scholars are using discussions of the system to explore what sociocultural assessment might look like in the context of university teaching (Campbell, personal communication).

Many of the tools for designing the assessments, as well as tools created in response to challenges linked to defining lesson objectives, are in widespread circulation among educators through the STEM Teaching Tools website housed at the University of Washington. Two of the most popular tools are the ones for supporting 3D assessment design, both of which have been downloaded more than 5,000 times. A third published last year called "Defining Meaningful Daily Learning Objectives," has been downloaded more than 1,000 times. Even though these locally developed tools are intended to address local needs, they clearly resonate with others more broadly.

References

Achieve. (2017). Using phenomena in NGSS-designed lessons and units. Washington, DC: Author.

Achieve. (2018). Criteria for procuring and evaluating high-quality and aligned summative science assessments. Washington, DC: Author.

Allen, C. D., & Penuel, W. R. (2015). Studying teachers' sensemaking to analyze teachers' responses to professional development focused on new standards *Journal of Teacher Education, 66*(2), 136-149.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, & Accountability, 21*(1), 5-31.

Boscardin, C. K., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and Algebra assessments. *Educational Assessment, 10*(4), 307-332.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation, 19*(2-3), 104-120.

Coburn, C. E., Penuel, W. R., & Geil, K. (2013). *Research-practice partnerships at the district level: A new strategy for leveraging research for educational improvement*. Berkeley, CA and Boulder, CO: University of California and University of Colorado.

Confrey, J. (2008). Framing effective and fair data use from high-stakes testing in its historical, legal, and technical context. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 33-54). New York: Teachers College Press.

Crouse, K., Gitomer, D. H., & Joyce, J. (2016). An analysis of the meaning and use of student learning objectives. . In K. K. Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice* (pp. 203-222). New York, NY: Macmillan.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Schank, P. (2010). Teaching routines to enhance collaboration using classroom network technology. In F. Pozzi & D. Persico (Eds.), *Techniques for fostering collaboration in online learning communities: Theoretical and practical perspectives* (pp. 222-244). Hershey, PA: IGI Global.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational quality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Heritage, M. (2018). Making assessment work for teachers. *Educational Measurement: Issues and Practice, 37*(1), 39-41.

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice*, 16-24.

Honig, M. I., & Hatch, T. C. (2004). Crafting coherence: How schools strategically manage multiple, external demands. *Educational Researcher, 33*(8), 16-30.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, dscovery,

problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Kolodner, J. L. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kolodner, J. L., Crismond, D., Gray, J., Holbrook, J., & Puntambekar, S. (1998). Learning by design from theory to practice. In A. Bruckman, M. Guzdial, J. L. Kolodner, & A. Ram (Eds.), *Proceedings of the International Conference of the Learning Sciences* (pp. 16-22). Atlanta, GA: ISLS.

Kolodner, J. L., Gray, J. T., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in Learning by Design classrooms. *Cognitive Science Quarterly, 3*(2), 119-170.

Krajcik, J. S., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education, 25*(2), 157-175.

Krajcik, J. S., & Mamlok-Naaman, R. (2006). Using driving questions to motivate and sustain student interest in learning science. In K. Tobin (Ed.), *Teaching and learning science: An encyclopedia* (pp. 317-327). Westport, CT: Greenwood Publishing.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27-48). Cambridge, UK: Cambridge University Press.

LeMahieu, P. G., & Reilly, E. C. (2004). Systems of coherence and resonance: Assessment for education and assessment of education. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (pp. 189-202). Chicago, IL: University of Chicago Press.

Lobato, J. (2003). How design experiments can inform a rethinking of transfer and vice versa. *Educational Researcher, 32*(1), 17-20.

Lobato, J. (2006). Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *The Journal of the Learning Sciences, 15*(4), 431-449.

McNeill, K. L., & Krajcik, J. S. (2012). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, reasoning framework for talk and writing*. Boston, MA: Pearson Education, Inc.

Nasir, N. i. S., & Cooks, J. (2009). Becoming a hurdler: How learning settings afford identities. *Anthropology and Education Quarterly, 40*(1), 41-61.

Nasir, N. i. S., & Hand, V. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research, 76*(4), 449-475.

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. (2001). *Knowing what students know*. Washington, DC: National Academies Press.

National Research Council. (2006). *Systems for state science assessment*. 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.

National Research Council. (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

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

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

Penuel, W. R., Bell, P., Neill, T., Shaw, S., Hopkins, M., & Farrell, C. C. (in press). Building a Networked Improvement Community to promote equitable, coherent systems of science education. *AASA Journal of Scholarship and Practice*.

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.

Penuel, W. R., Phillips, R. A., & Harris, C. J. (2014). Analysing curriculum implementation from integrity and actor-oriented perspectives. *Journal of Curriculum Studies, 46*(6), 751-777.

Penuel, W. R., & Reiser, B. J. (2018). Designing NGSS-aligned curriculum materials. Paper prepared for the Committee to Revise America's Lab Report. Washington, DC: National Academies of Science and Medicine.

Penuel, W. R., Van Horne, K., Severance, S., Quigley, D., & Sumner, T. (2016). Students' responses to curricular activities as indicator of coherence in project-based science. In C.-K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Proceedings of the 12th International Conference of the Learning Sciences* (Vol. 2, pp. 855-858). Singapore: International Society of the Learning Sciences.

Reiser, B. J., & Novak, M. (2017). *Developing coherent storylines of NGSS lessons*. Paper presented at the NSTA Area Conference, Milwaukee, WI.

Reiser, B. J., Novak, M., & McGill, T. A. W. (2017). *Coherence from the students' perspective: Why the vision of the Framework for K-12 Science Education requires more than simply "combining three dimensions of science learning*. Paper presented at the Board on Science Education Workshop "Instructional Materials for the Next Generation Science Standards, Washington, DC.

Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan, 98*(6), 47-52.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21-34.

Smithson, J. L., Porter, A. C., & Blank, R. K. (1995). Describing the enacted curriculum: Development and dissemination of opportunity to learn indicators in science education. Washington, DC: Council of Chief State School Officers.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137-156.

Weizman, A., Shwartz, Y., & Fortus, D. (2008). The driving question board: a visual organizer for project-based science. *The Science Teacher, 75*(8), 33-37.

Wilson, M., & Draney, K. (2004). Some links bewteen large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 132-152). Chicago, IL: National Society for the Study of Education.

Yeager, D., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). Practical measurement. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

EndNote

[1]The co-design process typically entails a process of rehearsing the anchoring phenomenon routine, which ensures that many—but definitely not all—of the student questions that make it onto the class DQB can be anticipated ahead of time with some reliability.