Running Head: PROXIMAL TRANSFER TASKS FOR 3D LEARNING

Developing Tasks to Assess Phenomenon-Based Science Learning:

Challenges and Lessons Learned from Building Proximal Transfer Tasks

William R. Penuel[1,2]

Michael L. Turner[2]

Jennifer K. Jacobs[1]

Katie Van Horne[1]

Tamara Sumner[1]

[1]*Institute of Cognitive Science* [2]*School of Education*

University of Colorado Boulder

Acknowledgments:

This material is based in part upon work supported by the Gordon and Betty Moore Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funder. The authors wish to thank the teachers who participated in this study and our district partners for their time and contributions to this study.

Abstract

The vision for science proficiency presented in *A Framework for K-12 Science Education* calls for significant shifts in both teaching and assessment. In this paper, we describe an effort to develop and validate a set of *proximal transfer tasks* for high school biology classrooms where teachers were implementing a problem-based curriculum. The proximal transfer tasks presented students with phenomena related to but distinct from the phenomena they had studied in class and asked students to apply their understanding of disciplinary core ideas, practices, and crosscutting concepts targeted in curriculum. We tested these tasks with a sample of 733 students from 11 teachers' classes. Each of these students completed two tasks prior to beginning the unit and two tasks after they had finished the unit. We found that nearly all pre- and post-test task pairs were aligned to written opportunities to learn in the curriculum, that is, students showed significant growth over the course of a unit. In addition, task pairs revealed differences across teachers. However, the relative growth of students depended on which tasks students completed, indicating wide variation in task difficulty. Our findings point to the potential of developing three-dimensional proximal transfer tasks but also to the difficulty of constructing equivalent tasks.

Key Words: assessment, transfer tasks, phenomenon-based learning, NGSS

Developing Tasks to Assess Phenomenon-Based Science Learning: Challenges and Lessons Learned from Building Proximal Transfer Tasks

The Next Generation Science Standards (NGSS; NGSS Lead States, 2013) present considerable challenges to assessment developers. Typical achievement tests in science assess knowledge of facts and procedures, but the NGSS represent science proficiency as evident when students can use science and engineering practices and apply their understanding of core ideas and crosscutting concepts in the context of explaining phenomena and solving problems (National Research Council, 2014; Pellegrino, 2013). Nearly all of the performance expectations of the NGSS are "three dimensional;" that is, they fuse or integrate an element of a disciplinary core idea, crosscutting concept, and science and engineering practice into a statement of what students know and can do.[1] Just how to design, score, and evaluate the quality of three-dimensional science assessments has been the focus of considerable effort and debate among researchers, assessment vendors, and local and state science education leaders.

Of particular interest and concern is how to develop assessments that can evaluate the quality of phenomenon-based teaching, a form of problem-based teaching that is anchored in helping students develop explanatory models of "observable events that occur in the universe and that we can use our science knowledge to explain or predict" (Achieve, 2017). Examples of phenomena include the evolution of bacteria that are resistant to antibiotics, the genetic causes of Duchenne Muscular Dystrophy (DMD), and the sudden increase then decrease in populations of large herbivores on the Serengeti.[2] Developing assessments of this type of teaching is key, because NGSS-aligned curriculum materials are expected to be organized around explaining phenomena or their engineering analogue, solving problems (Achieve, 2016). Further, anchoring curriculum and instruction in phenomena and design challenges is hypothesized to support the

development of students' integrated—rather than piecemeal—understanding of the three dimensions (National Academies of Sciences Engineering and Medicine, 2018b). Determining the accuracy of this hypothesis is a central challenge of the NGSS.

This paper presents results from a validity study for assessment tasks designed to evaluate phenomenon-based teaching as part of a high school life sciences unit on evolution. It draws on evidence from a large pilot study of a set of *transfer tasks* that are anchored in new, unfamiliar phenomena and that are intended to assess three-dimensional science proficiency of students who completed the unit to address the following questions:

- Can transfer tasks be designed that are aligned with opportunities to learn presented in curriculum materials?
- Can transfer tasks be designed that are fair for all genders and students of different racial backgrounds?
- How comparable are performances on phenomenon-based tasks intended to assess the same standards as the curriculum?

By addressing these three questions, we hope to add to our understanding of the possibilities and complexities of evaluating the efficacy of phenomenon-based teaching as a strategy for helping students master the performance expectations of the NGSS.

Challenges to Designing Three Dimensional Assessments

Typical achievement tests in science assess knowledge of facts and procedures, focused primarily on assessing performance at a single grade level. Most tests are comprised of independent multiple-choice items, each of which tests some aspect of science content (Alonzo & Ke, 2016; Blank & Adams, 2018). Until recently, there have been few examples of assessments that reflect a vision of science proficiency as the integration of understanding of

core ideas, practices, and crosscutting concepts (Harris, Krajcik, Pellegrino, & McElhaney, 2016; Wertheim et al., 2016). This is hardly surprising, as most state tests in use today reflect the focus of a previous generation of science standards, which emphasized primarily knowledge of facts and ability to carry out procedures isolated from specific content (Achieve, 2018a). To the degree that districts and schools are held to account performance on tests, assessments at the district and classroom level will likely reflect this image of learning in both form and content (Au, 2007). Until there are good models to follow, then, there is little incentive to develop three-dimensional assessments.

Developers of new assessments of integrated science proficiency face additional challenges, beyond contending with the limited number of good examples. The NGSS underspecify many aspects of what is to be assessed, leaving many decisions to developers as to how to structure assessment tasks in ways that elicit students' integrated understanding of science (Pellegrino, 2013). Further, guidance from standards underspecifies whether and how to integrate disciplinary core ideas, science and engineering practices, and crosscutting concepts within rubrics and score reports (DeBarger, Penuel, Harris, & Kennedy, 2015). Not surprisingly, analyses of assessments of the NGSS indicate that there are a range of "flavors" emerging for how to assess three-dimensional science learning (Achieve, 2018a; Alonzo & Ke, 2016).

Key Design Features of Three-Dimensional Assessments

Recently, staff from Achieve, Inc. (2018b) convened a group of assessment researchers and state leaders in science education to develop a set of criteria or guidelines for judging the quality of three-dimensional classroom assessments. In developing these guidelines, the group relied on multiple sources: recommendations developed by a committee of the National Academies of Science, Engineering, and Medicine (National Research Council, 2014), the ongoing research

and development efforts of several teams in different universities (e.g., Harris et al., 2016; Penuel, Frumin, Van Horne, & Jacobs, 2018; Wertheim, 2016), as well as the experiences of state science leaders in states that had adopted the NGSS. As such, these criteria reflect a rich blend of research-based and practice-based knowledge. Below, we describe these criteria and the justifications for their importance.

Tasks are driven by high-quality scenarios that focus on phenomena or problems. Assessing students' knowledge-in-use requires tasks in which students must apply their understanding of disciplinary core ideas and crosscutting concepts, using science and engineering practices (Pellegrino, 2013). If such tasks are to elicit students' ability to do science, moreover, such tasks need to engage students in multiple, "connected" practices (National Research Council, 2014, p. 130). Scientific practices are not isolated, but are integrated in service of answering specific questions to develop knowledge about the natural world (Kelly, 2012; Manz, 2015a). Further, demonstrating proficiency cannot be accomplished through multiple-choice items alone, since these do not require students to actively construct and integrate knowledge through engaging in practices in the ways called for in the NGSS (Lee, Liu, & Linn, 2011; National Research Council, 2014, p. 6; Nehm, Beggrow, Opfer, & Ha, 2012). The NRC (2014) committee concluded that the best way to assess the three dimensions simultaneously was through what it termed "multicomponent" assessment tasks, that is, tasks comprised of multiple related prompts organized around a single problem context, or what we call here a "scenario." Its reasoning was that such tasks presented multiple pieces of evidence that, when considered as a whole, could provide "a sufficient indicator of student understanding" (National Research Council, 2014, p. 89). Further, organizing clusters of questions around a single phenomenon to be explained (using science practices) or problem to be solved (using

engineering practices) provides an ideal context for eliciting how well students' reasoning and sensemaking approximates certain aspects of how scientists and engineers really work (Manz, 2015a, 2015b).

Tasks require sense-making using the three dimensions. Even with tasks that are organized around scenarios that present phenomena to be explained or problems to be solved, there is the risk that tasks can be completed by rote in ways that undercut rather than support the vision behind the NGSS (Alonzo & Ke, 2016). For example, it is relatively easy to write multicomponent assessment tasks that are comprised of prompts that are disconnected from the overarching scenario and from one another. To avoid this risk, tasks need to prompt students to engage students in sensemaking, that is, in a process of building an explanation that resolves a gap or conflict in knowledge (Odden & Russ, 2018) or designing a solution that addresses an identified individual, community, or socio-ecological need (Gunckel & Tolbert, 2018). Beyond being organized around phenomena and problems, then, assessment tasks must be organized to demand that students use their understanding of core ideas, practices, and crosscutting concepts to make sense of the scenario presented to them. This will likely require that the scenario point explicitly to gaps in knowledge or needs to be addressed. This goal may be accomplished through an explicit problematizing of the context, to highlight disciplinary ideas at stake and to motivate students (Engle, 2012; Reiser, 2004). The sequence of prompts will need to be intentionally designed, so as to elicit relevant student resources for sensemaking as starting points, and building incrementally toward more complete explanations or solutions, much like the expectations of curriculum designed to align with the NGSS (National Academies of Sciences Engineering and Medicine, 2018b). Finally, each prompt should be designed to be *in the service of* sensemaking about phenomena and problems, such that knowledge of the

dimensions is not elicited as isolable knowledge or skills in a manner that would undermine the image of science learning presented in the *Framework for K-12 Science Education* (National Research Council, 2012).

Tasks are fair and equitable. An important consideration in determining whether a test is fair and equitable is whether it is accessible. Accessibility refers to whether assessments provide a way for all students to participate that yields information about each student's proficiency that is valid for the purposes for which it was intended to be used (Thurlow & Kopriva, 2015; Thurlow et al., 2009). Ensuring access in some cases may require assessment accommodations, in which materials and administration protocols are altered for students (e.g., those with particular identified disabilities) to enhance their ability to respond to tasks without altering the learning goals to be assessed (Lazarus, Thurlow, Lail, & Christensen, 2009; Thurlow, 2012). Another strategy used in assessment design is to employ the Universal Design for Learning Framework (Rose & Meyer, 2002), which demands different ways to present assessment materials and allows students to use different ways to engage with tasks and express what they know. A third strategy for making assessments more accessible is to provide some scaffolding that guides students as to the depth or scope of a response that is required (Abell & Siegel, 2011; Kang, Thompson, & Windschitl, 2014; Siegel & Wissehr, 2011; Underwood, Posey, Herrington, Carmel, & Cooper, 2018). For English learners, using non-textual elements (e.g., visuals) to communicate essential meanings in assessment materials can be a useful scaffold to enhance access to tasks, as can presenting scenarios in students' native language (Buxton et al., 2014; Pennock-Roman & Rivera, 2011; Siegel & Wissehr, 2011).

Tasks that are fair and equitable are ones in which students taking the assessment can find meaning in relation to their interests, everyday experiences, and identities. This aspect of equity

is specifically called out as a key assumption of *A Framework for K-12 Science Education* (National Research Council, 2012) and is based on a body of evidence cited in the *Framework* that points to the importance of connecting instruction to students' interests and experiences. Similarly, instruction that is relevant to students' lives promotes equity because it helps them see how science and engineering can be useful to making their communities better places to live (Penuel, 2014; Tan & Calabrese Barton, 2012). Designing assessment tasks that connect to students' life experiences requires knowledge of students' interests, identities, and cultural communities (National Research Council, 2014; Shepard, Penuel, & Pellegrino, 2018). It also requires analysis of assessment results for bias that might result from some contexts or terms being so unfamiliar to particular groups of students that they cannot access the task.

To be fair and equitable, tasks need to assess learning goals that students have had sufficient opportunities to learn (National Research Council, 2014). When students encounter tasks for which they have not been prepared, either because instruction has not focused on the learning goals being assessed or because it was not of sufficient quality or duration, then valid interpretation of results is compromised. Tasks should be designed with consideration for the curriculum that students will encounter, as well as the instructional approaches used (Shepard, 2009).

Tasks support their intended targets and purpose. To conclude that a task meets its intended targets and purposes, assessment designers need clear answers to questions about the claims they want to be able to make about assessments, how best to elicit artifacts of student thinking that can provide data for those claims, and guidance to interpreters for how to use that data to evaluate what students know and can do (Messick, 1994; Pellegrino, 2013). Such a process necessarily begins with defining and elaborating the constructs to be assessed (Federer, Nehm,

Opfer, & Pearl, 2015; Mislevy & Haertel, 2006). In science, assessment designers may undertake detailed analyses of performance expectations that might be taught across an entire unit (Achieve, 2015), specifications of hypothetical learning progressions (Mohan, Chen, & Anderson, 2009), or the construction of finer-grained three-dimensional learning goals appropriate for assessing shorter chunks of instruction (Harris et al., 2016). The tasks that are designed should reflect the analysis of the constructs, that is, they should elicit student work artifacts that are relevant to the constructs and representative of the different facets of the construct that designers identified (Messick, 1995). At the same time, they should avoid bias that could arise from some students encountering tasks that rely on background knowledge not relevant to mastery of assessed learning goals (Messick, 1995). Answer keys or scoring guides, in turn, should help interpreters of student artifacts fairly judge the degree to which students have mastered the learning goals to be assessed (Abedi & Gándara, 2006).

Developing validity evidence presents different kinds of challenges to developers, depending on the intended use of assessments. Districts and states use assessment evidence primarily for monitoring purposes; some states will incorporate such evidence into accountability scores for schools as well, making those tests high-stakes (Achieve, 2019). Assessments used for monitoring and accountability purposes must be shown to be reliable, fair, and efficient to administer; the demands that students demonstrate their ability to apply science knowledge makes all three of these goals difficult to obtain (National Research Council, 2014). Teachers use classroom assessment data for different purposes, including for grading students and for adjusting their instruction when they discover students are experiencing difficulty mastering particular learning goals. The validity of classroom-based formative assessments relies on evidence that they can be used effectively to improve student learning (Pellegrino, DiBello, &

Goldman, 2016). Such assessments are time-consuming to develop and test, to the degree that they require richly elaborated theories of learning and evidence regarding different means to accomplish student learning goals (Penuel & Shepard, 2016). Assessments used to evaluate instructional approaches or curriculum materials—the type that is the focus of this paper—require evidence related to fairness and reliability, as well as to sensitivity to opportunities to learn provided in curriculum. That is, tests must be able to detect growth from learning through particular instructional methods or with particular sets of curriculum materials. These assessments might be used by evaluators to judge the merit or worth of materials, or—as in our case—by designers of curriculum materials who are seeking to improve those materials.

A good example of an effort to develop reliable, instructionally sensitive, and fair assessments for evaluation purposes is that of DeBarger and colleagues' evaluation of the Project-Based Inquiry Science (PBIS; Kolodner et al., 2008) curriculum. The team used an evidence-centered design process to develop assessments that integrated two of the three dimensions of the NGSS, science and engineering practices and disciplinary core ideas. The team developed, tested, and revised their assessments over the course of two years, developing evidence of reliability, then testing a revised version in a field trial (DeBarger et al., 2015). The field trial showed that the assessments were sensitive to instruction and showed no bias against or for any particular group with respect to student growth (Penuel et al., 2015). As the assessments developed for the current study do, the item clusters were focused on asking students to develop explanatory models of phenomena. The current effort, however, is distinct in that the multicomponent tasks sought to assess all three dimensions, and the evaluation was focused on an approach to teaching that was more thoroughly anchored in phenomena.

What Should We Attend to in Assessing Phenomenon-Based Learning

A key consequence of the integration of the three dimensions in the NGSS is that learning requires a meaningful purpose for using science and engineering practices to work with science ideas. The *Framework* (National Research Council, 2012, p. 50). defines these purposes: "Science begins with a question about a phenomenon… and seeks to develop theories that can provide explanatory answers to such questions." Analogously, "Engineering begins with a problem, need, or desire that suggests an engineering problem that needs to be solved." Thus, integration of the *Framework's* three dimensions means more than simply focusing students' attention at some point in a lesson on each of the three dimensions in separate parts of the work. Explaining phenomena and solving problems are key to integration.

In some models of instruction for helping students master the performance expectations of the NGSS, phenomena or engineering problems serve as anchors for coherent sequences of lessons (National Academies of Sciences Engineering and Medicine, 2018b). By "anchor," we mean that phenomena and problems provide the primary context for students to ask questions, around which a sequence of investigations can be organized that help students build understanding incrementally of an observable event in the world (Schwarz, Passmore, & Reiser, 2017). In teaching that is phenomenon-based, phenomena are more than just initial "hooks" to capture student interest, they are what motivates the need for developing understanding of science ideas and what serves as a "spine" that links lessons together into a coherent whole (Penuel & Reiser, 2018). At the same time, phenomenon-based learning does prize student interest and experience, as well as students' different perspectives on phenomena being studied (Francis, Breland, Østergaard, Lieblein, & Morse, 2013; Symeonidis & Schwarz, 2016). Therefore, successful phenomena are ones that are compelling to students, that is, they both

attract and sustain students' attention and engagement through the course of a unit (Blumenfeld, Soloway, Marx, Guzdial, & Palincsar, 1991).

Neither student engagement nor learning from phenomenon-based learning can be assumed, and past research shows mixed evidence of success of learning anchored or based in specific problems. If we treat phenomenon-based learning as a form of learning that shares many features with or as a type of problem-based learning, then lessons learned about problem-based learning are relevant to the challenge of assessing phenomenon-based learning. Reviews of problem-based learning across a wide variety of fields show generally positive outcomes, but also considerable heterogeneity (Hmelo-Silver, 2004; Walker & Leary, 2009). Chief among the benefits cited for problem-based learning are flexible application of knowledge to new settings, and increase in problem-solving skill (Cognition and Technology Group at Vanderbilt, 1997; Hmelo-Silver, 2004; Kolodner, 1993). Problem-based learning can also result in increased knowledge integration, when students are able to tie general ideas and concepts to the specific problem they are solving (Lu, Bridges, & Hmelo-Silver, 2014). Major causes of variability in outcomes are not well understood, but significant enough to warrant empirical investigation with carefully designed measures appropriate to the task of measuring the impact of different forms of problem-based learning, including phenomenon-based learning (Walker & Leary).

The specific challenges that qualitative researchers have identified as linked to problem-based learning provide some additional guidance for the design of assessments that could be used to evaluate phenomenon-based learning. One is the problem of generalization: when students' learning is anchored in specific problems, they may not generalize from problems to broader principles or ideas (Kolodner, 1993; Lehrer, Schauble, & Petrosino, 2001; Petrosino, 1998). Therefore, if we expect students to gain mastery of core ideas and crosscutting concepts through

phenomenon-based teaching, then we need to assess whether they can apply those core ideas and crosscutting concepts in problem contexts that must be carefully chosen to reflect what students have had the opportunity to learn but that are new and unfamiliar. We will also need to assess their grasp of science and engineering practices in this new context, ideally assessing how they engage in practices targeted in a given performance expectation. If mastery of a practice involves not just knowing how to use a practice but also when to engage in it within ongoing activity (Ford, 2015; Manz, 2015a), then assessments are needed that can elicit students' knowledge of how to tackle the new problem without giving away the procedures for doing so.

Following Ruiz-Primo, Shavelson, Hamilton, and Klein (2002), we refer to tasks with novel contexts for students to apply learning from phenomenon-based learning as *proximal transfer tasks.* They are proximal to curriculum, in that they are intended to assess mastery of the content (i.e., NGSS performance expectations) that is taught in specific sets of materials at the unit level. Here, proximal contrasts to distal tasks, which are tasks that assess different content or content at the annual or grade-band level, as is common in state accountability tasks. The tasks we designed for this study were "transfer" tasks, in that the situation or phenomenon that they are presented is not one they have seen before. The tasks did not provide students with the science ideas they would need to answer prompts given to them. Further, although they were given some scaffolding with respect to the practices to use to develop an explanation of the phenomenon presented in the transfer task, students had to use practices to make sense of phenomena presented to them in writing.

The Current Study

In the current study, we set out to develop validity evidence for proximal transfer tasks that meet the criteria for assessment of three-dimensional science learning embodied in the National

Research Council's (2014) report, *Developing Assessments for the Next Generation Science Standards,* and as reflected in guidance by Achieve, Inc. Specifically, we set out to address three research questions:

- Can transfer tasks be designed that are aligned with opportunities to learn presented in curriculum materials?
- Can transfer tasks be designed that are fair for all genders and students of different racial backgrounds?
- How comparable are performances on different phenomenon-based tasks intended to assess the same standards as the curriculum?

Below, we describe our sample for pilot study, the tasks we developed and process for developing them, methods for piloting the tasks, and our approach to scoring and analyzing data from the pilot study. The data presented are from our team's second major pilot study of the assessments, and the tasks tested here reflect lessons learned from earlier versions of the assessments and associated scoring guides.

Sample

Our sample was comprised of students from 11 different high school teachers' biology classrooms, across five high schools. The teachers came from a large, racially diverse, urban school district in the U.S. Midwest. The student demographics were 55.5% Hispanic, 23.2% White, 13.4% African-American, 4% other, 3.2% Asian, and 0.6% American Indian, Over two-thirds (68.5%) of the students received free or reduced price lunch, and over one-third (36.8%) were English Learners. All students tested were in biology classrooms during Fall 2017 that were

implementing a unit on evolution, *Why Don't Antibiotics Work Like They Used To?* The curriculum materials are available online at http://tinyurl.com/ihubEvoLandingPage.

The district had adopted the Next Generation Science Standards, and their teachers had received extensive professional development in the standards and in phenomenon-based teaching in science. Teachers could choose to participate in multiple professional learning experiences led by members of the research team and/or district science administrators including full-day workshops, after-school meetings, professional learning community meetings, and individualized coaching. All of the teachers included in this study took part in at least some of these professional development options.

Our initial sample was comprised of a total of 733 students from 11 teachers' classes. Each of these students completed pretests prior to beginning the unit and posttests after they had finished the unit. All of the tests contained two tasks, and students did not get the same tasks from pre to posttest. For all tests we decided to remove students from the sample who made no attempt to complete at least one of the two tasks, as such data would be insufficient to evaluate the comparability of tasks. While the absence of an attempt to complete a task could provide evidence of a task's difficulty, the majority (64%) of students removed from the sample did not provide any responses for the second of two tasks given on the assessment, suggesting that they did not have time to attempt the task. Four students left both tasks blank on a test and were also removed from the sample. This reduced the sample of students from 733 to 583 for conducting analysis. Table 1 below shows the gender and ethnic composition of the final analysis sample.

Insert Table 1 about here

Evidence Related to Implementation

We have some evidence related to curriculum implementation for a little over half (n = 6) of the classrooms in the study, which we present to illustrate that—for these classrooms at least—students did have opportunities to engage with the curricular content and activity formats to which the assessment tasks were aligned. This evidence comes from student surveys teachers administered to their classes between one and four times during the teaching of the unit. These student surveys focus on three important aspects of the curriculum, namely its coherence from the student point of view, student contributions through discussion, and the perceived relevance of classroom activities. As Table 2 shows, for the teachers for whom we have implementation data, the implementations were, overall, consistent with the intent of the curriculum from the standpoint of student experience. A high percent of students say that they made progress on questions from the class "Driving Question Board," which is a record of the questions that students generate at the beginning of the unit, in order to explain the anchoring phenomenon. In addition, nearly all students say they contributed to discussions in class, with a significant percentage participating in both small group and whole class discussions. And nearly all said the lesson had some relevance to them personally, the class, or the community.

Insert Table 2 about here

Task Development and Descriptions

Each of the tasks was developed to serve as a test of transfer of understanding for the focal unit that was aligned to one or more performance expectations of the NGSS. The goal was to produce multi-component assessment tasks that were proximal to the instruction they received, were organized around a real scientific phenomenon, and included real scientific data related to that phenomenon. Proximal assessments include tasks in which the focal standards—or performance expectations, in the case of the NGSS—are the same as those taught in class, but where the context is unfamiliar (Ruiz-Primo et al., 2002). The assessments presented to students were intended to assess their understanding of five performance expectations in high school life science related to biological evolution. At various points in the unit the curriculum provides students with opportunities to develop understandings of each of the assessed performance expectation. The contexts in which they investigated evolution were two "anchoring" phenomena: (1) the evolution of antibiotic-resistant populations of bacteria and (2) the microevolution of a population of dark-eyed juncos in response to migration to a new environment for breeding. The assessments present students with different phenomena to explain (see Table 2 below) and invite them to consider how what they learned in the unit can help them make sense of and explain these new phenomena.

Domain analysis. Assessment development began with an analysis of the domain by the team that developed the curriculum units, which included one member of the committee that developed *A Framework for K-12 Science Education* (National Research Council, 2012). As a

whole team, the group conducted a systematic "unpacking" or analysis of all text related to the *Framework* for each of the performance expectations to be assessed. The unpacking was comprised of sub-claims developed for each sentence in the Framework related to DCIs, as shown in Figure 1 below.

Insert Figure 1 about here

For the Science and Engineering Practices (SEPs) and Crosscutting Concepts (CCCs), the team relied on domain analyses produced as part of the development of the NGSS and included in Appendices F and G. Appendix F of the NGSS outlines specific components of all eight of the SEPs for each grade band. We used the components identified in the NGSS to develop prompts related to the SEPs. For the CCCs, we used Appendix G's guidance related to grade-band targets in a similar manner as we used Appendix F for the SEPs.

We also relied on information included in Clarification Statements, that is, statements embedded in the standards to clarify their intent, and Assessment Boundaries, statements about what should and should not be assessed with respect to the performance expectation, when considering how to focus our assessments and, as noted below, in our specification of task requirements.

Developing task specifications. We set out to design tasks that adhered to broad guidelines presented in the National Research Council (2014) volume, *Developing Assessments of the Next Generation Science Standards.* As such, each task involved multiple components or prompts aimed at eliciting information related to each of the three dimensions of science proficiency. We did not seek, however, to develop claims related to each of the dimensions separately or regarding any sub-claims we identified as part of the domain analysis. Rather, our intention was

to support claims about student proficiency at the level of the performance expectation, that is, their integrated science understanding. We set out to design tasks that were organized around a single scenario, in which students would be presented with a textual description of a phenomenon to be explained and some evidence related to the phenomenon.

We used our analysis or unpacking of each of the three dimensions to first establish a set of criteria for selecting phenomena that could present students with an opportunity to apply their understanding of Disciplinary Core Ideas (DCIs). For all tasks, we sought to identify a phenomenon that required students to engage with evidence from an actual scientific study conducted in the past 50 years. A second criterion applicable to all tasks was the availability of actual data in the form of observations or experiments that were related to the DCI. Rather than have students construct their own data for tasks (as is called for in some performance expectations), we constrained our assessments to ones in which students would be asked to analyze data and, in some cases, predict what might happen in the future based on evidence presented in the task and their understanding of a DCI. For example, a phenomenon related to speciation and common ancestry would need to be accompanied by data related to the determination of genetic similarity of two populations of organisms, while a phenomenon that pertained to adaptation and natural selection would need to be accompanied by data related to changes in distribution of traits in a population, as well as data related to how the environment was changing.

For each candidate phenomenon, we first constructed a scientific explanation that we imagined a proficient high school student might write as an explanation for the phenomenon. We used this explanation to help clarify ways that an understanding of the knowledge components for each of the three proficiency dimensions might be required to explain the phenomenon. In

addition, we used it to develop specific prompts (questions or items within tasks) based on the design team's intuitions regarding whether students would be likely to include a particular knowledge component in their explanation. These discussions involved much debate about how much scaffolding to provide; a debate that could not easily be resolved without evidence from student responses to alternate forms of each task. In the end, the nature and level of scaffolds in tasks we designed is best described as a *variable task feature* (Mislevy & Haertel, 2006). Variable task features are aspects of the assessment that can be varied in order to shift difficulty or focus.

We also used our analysis of the domain to define how to present evidence to students, such that the tasks require students to make sense of data within the boundaries of assessment appropriate to high school students. For example, a candidate study involved an evaluation of the similarity of populations' alleles that presented results from a Bayesian statistical analysis (Kearns et al., 2016). However, most high school students are not exposed to these kinds of statistics, and the NGSS boundary statements are explicit about limiting assessment to basic concepts of statistics and probability when assessing the SEP of Using Mathematics and Computational Thinking.

Task Development. We developed tasks through an iterative process, in which we identified and selected candidate phenomena, wrote tasks that included scenarios that represented the phenomenon to students and prompts for them to answer, developed hypothetical student responses to individual prompts, and created scoring guides. The process of iteration involved team-level review and, for each of the assessment tasks included in this analysis, pilot testing of assessments with students and reviewing scoring guides with expert scorers. Below, we describe

the final four tasks presented to students for this study. Table 2 below presents a summary of the descriptions.

Insert Table 3 about here

Evolution of Swallows. This task is based on a scientific study reported in *Current Biology* (Brown & Brown, 2013) describing the microevolution of a population of swallows after the construction of an interstate highway in western Nebraska. The swallows used highway overpasses as nests, and those swallows with longer wings had difficulty making quick evasive maneuvers to escape from cars when they retrieved food from the highway. Scientists documented decreases in average wing length over time, and as it did, the number of road kills decreased, indicative of natural selection and adaptation. The task is intended to elicit students' mastery of HS-LS4-3, which focuses on students' integrated understanding of natural selection and adaptation, analyzing and interpreting data, and patterns. It is also intended to elicit mastery of parts of LS-HS4-4, which focuses on students' integrated understanding of adaptation, explanation, and cause and effect (the distinction between causality and correlation is not assessed). In the task, students use information provided in the scenario to characterize advantages and disadvantages of nesting under highway overpasses. They are asked to draw inferences about population size by interpreting patterns in data related to roadkill and nests over time. Students are asked to develop conjectures about the causes of changes in the swallow population, and then use data presented to support claims about competitive advantage of shorter wings for swallows in the environment. To get full credit, they must apply the concept of natural selection within an explanation of what is happening in the population of swallows. Finally,

students are asked to make a prediction about what will happen to bird wing length in the future, using patterns observed in the data.

Galápagos Ground Finches. This task is based on studies conducted during an extended multi-year drought on Daphne Major, one of the Galapagos Islands, and it focuses on species known to Darwin, ground finches (Boag & Grant, 1981; Gibbs & Grant, 1987). Over time, in response to the drought, differing populations of finches evolved with respect to wing length and beak length. The graphs of data presented to students in the task come from another learning sciences research project (Reiser et al., 2001), rather than from the original studies. Like the Swalllows task, the Finches task is intended to elicit students' mastery of HS-LS4-3 as well as parts of LS-HS4-4. In the task, students are asked to identify patterns in graphs related to proportions of finches with different wing and beak lengths over time. Next, they are prompted to construct an explanation relating changing environmental conditions and changes in the proportions of traits in a population, using the concept of competitive advantage as an explanatory principle. Students are asked to then make a prediction about what will happen to proportions of birds with longer beaks if conditions stay the same in their environment. Finally, they are asked to construct an explanation for how environmental conditions must have changed, given a change in proportion of birds with long wings and knowledge that long wings allow birds to fly farther for food.

Two Species or One? This task is based on research findings presented in the journal *Conservation Genetics* and pertains to a study conducted to determine whether a population of robins in the Southwest Pacific Islands constitute a distinct species (Kearns et al., 2016). The issue is salient to society because the islands have undergone significant habitat loss and climate change, threatening the survival of the robins. The task is intended to elicit students' mastery of

HS-LS4-1, which focuses on students' integrated understanding of evidence of common ancestry and diversity, obtaining, evaluation, and communicating scientific information, and patterns. It also is intended to assess most elements of HS-LS4-5, which focuses on students' integrated understanding of adaptation, engaging in argument from evidence, and cause and effect (the distinction between causality and correlation is not assessed). In the task, students begin by analyzing patterns in genetic similarity data to draw conclusions about which birds have most recent common ancestor. They are then tasked with constructing an explanation for whether the birds are a distinct species, using evidence presented and knowledge of how scientists determine species membership. Last, they are asked to make a prediction about what will happen to two different populations of birds whose environments change and who do not interbreed.

Human Adaptation on the Tibetan Plateau. This task presents a phenomenon studied by Cynthia Beall and colleagues (Beall, 2007; Beall, Song, Elston, & Goldstein, 2004), notably the adaptation of human populations to living at high altitudes on the Tibetan Plateau. Students are presented with the challenge of explaining why visitors' physiological response to high altitude is different from the response of human beings who have lived there for many generations. Like the Species task, the Tibet task is intended to elicit students' mastery of most elements of HS-LS4-5. In the task, students are asked to use the given data to construct an explanation for why—over many generations--Tibetans living at high elevations are physiologically different from humans living at lower elevations. In addition, they are asked to use evidence to support the claim that change in allele distribution is due to adaptation. Finally, students make predictions about what the frequency of different characteristics and alleles might be in a Tibetan population based of their analysis of the data and what they know about adaptation.

Test Construction and Assessment Administration. Piloting of assessments indicated that we could present students with no more than two of the four tasks per class period. Therefore, we created test booklets comprised of different combinations of two tasks (Table 4). We instructed teachers to distribute all six forms within their classrooms at random, at the time of pretest. For the posttests, students received a test with two new tasks that they had not seen.

This particular design provided us with two advantages for addressing our research questions. First, the design enabled pairwise comparisons between tasks with respect to students' performance. Between task, within person differences in performance could be analyzed on posttest scores and for gains between pre- and post-test scores. Second, distributing all of the tasks within a single classroom provided a means to "unconfound" student growth linked to teacher effects from the differences in student growth that might be attributed to task difficulty. Table 5 below shows what tasks appeared on what forms, as well as the point value for each task and test form.

This test design has some limitations as well. By design, a student does not see a task twice, so no direct assessment of growth was possible. This also meant we have no "linking items" for establishing equivalency, a decision that proved consequential, given our findings.

Insert Tables 4 and 5 about here

Scoring of Assessments. We scored each of the assessment prompts using a rubric that assigned points based on identifying facets of understanding that should occur in "ideal" constructed responses. For example, when asked to explain why Tibetans living in the mountains differ physiologically from people visiting the mountains, students can earn up to four points if they (1) accurately describe a trait in the population, (2) connect the trait to survival, (3) connect

survival to reproduction, and (4) discuss natural section or adaptation. Table 6 provides a more detailed look at the scoring of this question, including required components and examples for each of the four points. The scoring guide was constructed through an iterative process that unfolded over the first two years of the project, guided first by hypothetical student responses and then by evidence from actual student responses.

Insert Table 6 about here

Two coders, both of whom had been involved in developing the scoring guides, completed scoring of all student tasks using the following procedure. First, they scored a set of sample student responses for each task, across multiple teachers, together as a team. Next, to establish reliability, they independently scored a different set of responses from at least 10 students for each task prompt. The target was to achieve ≥80% inter-rater agreement within each pair of coders for each prompt. If after a test of inter-coder agreement, the team did not meet this target, they revised the coding guide on the basis of discussion of discrepancies and then repeated the reliability test. This process continued until the team determined no further improvements to intercoder agreement could be achieved. Ultimately, the team of coders was able to achieve ≥80% agreement for each prompt, between 85-95% agreement for each task, and 90% agreement overall across the four tasks.

In addition, the two coders established midpoint reliability after they applied the scoring guides to approximately half the data. Again, the goal was to achieve ≥80% inter-rater agreement on each task prompt. However, on 5 of the 20 task prompts the raters fell short of this goal. In these cases, the coders discussed their differences and made necessary revisions to the scoring guides. They then attempted to establish midpoint reliability again, and after reaching ≥80%

agreement on each prompt the coders went back through all of the assessments and updated the scores as necessary to match the revised scoring guides.

Analysis of Results. We first summed scores by task, and then converted all scores into a percentage correct for each test form, since there was not consistency in the number of points awarded per task. Table 4 shows the total possible number of points per test form, from which percent correct scores were derived. Total scores ranged from 26 to 28 possible points, with four of the six forms having a total max of 27 points.

As an initial estimate the sensitivity of the tasks to instruction taken as a whole, we used the following formula:

Mean % Correct Posttest – Mean % Correct Pretest / SD pooled

Note that in this formula, we did not control for any observed differences in task difficulty or for teacher effects in this analysis.

To analyze differences in scores attributable to test form and teacher, we fit a linear regression model to the data. The outcome of interest was gain score, calculated as a percent correct on posttest minus percent correct on pretest. Predictors in the regression equation were teacher and test form.

Results

We found that the evidence regarding task comparability was mixed. For instance, Table 7 displays the descriptive data for each of the four tasks according to whether they were administered before instruction ("Pre"), or after instruction ("Post"). On the one hand, the baseline or pre-task percentage of points earned fell within a relatively narrow range (28-34%) for three of the four tasks. For the Finches task, baseline scores were much lower, between 19-23% on average, suggesting this task was considerably more difficult than the other three at

baseline. An analysis of post-test scores further confirmed that the Finches task was the most challenging for students, even after instruction.

Insert Table 7 about here

In addition, Table 8 indicates that the order in which tasks were presented to students on the assessments influenced their performance. A direct comparison is possible with two of the tasks, Finches and Swallows, which were presented to some students as the first task (Finch1, Swallows1) and to other students as the second task (Finch2, Swallows2). On both pre and post administrations, students performed more poorly on tasks presented second, suggesting test fatigue and/or lack of time to complete the assessments were important factors.

Insert Table 8 about here

Table 8 displays basic descriptive information for each of the eleven teachers who participated in the study as a function of students' improvement from pre to post tasks. For the remainder of the paper we will refer to the change between pre task and post task as the "gain score" even when there is a decrease in the scores. Table 7 shows that most of the participating teachers had similar results when all task combinations were aggregated at the teacher level. The major exception to this trend is Teacher F, whose students demonstrated significantly higher gains on the average relative to the students of other teachers. Both the minimum gain score and maximum gain score in Teacher F's class were almost a standard deviation higher than other participating classrooms. The view provided in Tables 6 and 7, which aggregate all of the task combinations, tends to mask the variability in score change, which we believe is worthy of further examination.

Insert Figure 2 about here

As a means of displaying how score change occurred between each combination of tasks (described in Table 4), Figure 2 presents a boxplot of the score distribution for each possible combination of pre-task and post-task. One might think of the individual columns of the box plot as an individual graph showing the distribution of gain scores for that columns' set of tasks. For instance, the rectangular box in the column labeled "Finch1-Swallows2" conveys that the middle 50% of students who took that combination of tasks had a gain score between -5% to +30%. The thick black bar in the upper portion of that box represents the median gain score, which in this case is 20%. In other words, the median student of all those who received these two tasks showed a 20% improvement from Finch1 to Swallows2. These students would have received assessments B2 (pre) and B1 (post). On the pretest, Finches was the first task (Finches1) and on the posttest, Swallows (Swallows2) was the second task. The lines coming out of the top and bottom of the box represent the upper and lower 25% of students in the gain score distribution. In some cases, individual values, represented by dots, extend out of the top of bottom of each plot. These points represent outlier values.

To provide an alternative view of the pre-post task differences, Table 9 displays the average gain for each of the test pairs along with the standard deviation of gain score. In terms of gains—an indicator of differential sensitivity to instruction—these ranged widely, with pairing of Tibet2 followed by Finch2 showing a negative gain of 5 percentage points, and at the other extreme, Finch2 followed by Tibet2 showing a positive gain of 29 percentage points. If all pairs involving the most difficult task, Finches, are removed the range is restricted somewhat to an average gain

of 4 to 18 percentage points. This still remains a significant gap, suggesting that pairings of these tasks mattered for student performance.

From another perspective, with only a small number of exceptions our pairs of performance tasks appear sensitive to opportunities to learn from the curriculum materials. In 13 of 15 pairs of pre- and post-tasks students showed an improvement on the percentage of points earned on the tasks. Given the tasks' focus on performance expectations targeted in the phenomenon-based unit, this result is encouraging. If the most difficult task (which we have deemed to be the Finches task) was removed from the analysis, then all pairs would yield improvements from pre- and post-test. Moreover, the gain from pre- to post-test would be approximately +0.48 standard deviations.

Insert Table 9 about here

To further explore how the combinations of tasks compared, we ran a linear regression that regressed task combination on gain score. These results, displayed in left columns of Table 10, show the regression coefficients for each of the task combinations with the exception of the Tibet2-Finch2 task combination, which was used as the base-line dummy variable and comparison point for all of the other task combinations. As Table 9 demonstrates, only two task combinations were not significantly different from each other: Tibet2-Finch2 and Swallows1-Finch2. Unsurprisingly, these are the only two task combinations that contained median and mean test scores that showed no improvement between pre and post task. These results also suggest that many of the gain scores from the task combinations are significantly different from

each other, providing even more evidence that task combination affected student performance in meaningful ways.

Insert Table 10 about here

As a final set of analysis, we examined whether the addition of teacher to the task combination regression equation would significantly alter our results. The inclusion of this extra variable accounts for variability across teachers in how materials were taught, how students interacted, and other differences between groups of students. As the middle columns in Table 9 indicate, teacher was a significant variable, but its inclusion did not alter the results for task combination. In other words, while teacher variability did affect student performance on the task combinations, each task combination, with the exception of Tibet2-Finch2 and Swallows1-Finch2, continued to display significantly different results. These results corroborate the previous suggestions that task combination affects performance.

We also included Gender and Ethnicity variables in the regression, which are included in the rightmost columns of Table 10. Our analyses indicate that gender and ethnicity are not significant factors, with the exception of the Hispanic ethnicity category. This category saw a small but significant decrease in test scores when compared with the White reference category. Thus, Hispanic students scored slightly (about 3%) lower than their peers. Similar to the inclusion of Teacher in the regression, these results do not alter the previous regression results and do not significantly alter the R^2 values, which suggests they explain very little variability in our model.

While the teacher effects do not significantly alter the variability in task combination, in some ways we find their statistical significance encouraging. In particular, we would expect teachers to implement curricular materials in different ways based on their own personal teaching approach in concert with their perception of their students' unique needs. Thus, any set of task combinations should be able to detect teacher effects and measure how teachers vary in how they implement the same materials. With that said, the low R^2 values (0.12-0.20) suggest that much of the across teacher variability might not be captured by the present set of tasks, which suggests that further revisions are warranted.

Analyses of Prompts in the Finches and Swallows Task

To help understand why performance might differ between tasks, we next turn to an interpretive analysis of task differences, using our own analyses of task qualities. A comparison of performance by prompt on two of the tasks that are most similar in terms of their performance targets and yet most distinctive in the overall performance of students helps illuminate potential problems with specific task prompts. The prompt that focused on eliciting students' understanding of the crosscutting concept of patterns proved much harder in the Finches task than in the Swallows task, even though both asked students to consider patterns regarding two different variables. It was not clear to our team as to why this prompt was more difficult, since the graph interpretation was more complex in the Swallows: students had to interpret data where there were two y-axes, not just one.

For both tasks, task performance was low when students had to supply the mechanism without scaffolding, in terms of passing on advantageous traits to offspring, who themselves were more likely to survive to reproduce. It may be that both tasks require more prompting to address survival advantage. Alternately, the questions could be made into multiple-choice

questions, which could also constrain the possible answer space for students; however, such a transformation would limit students' opportunities to show their application of understanding of the disciplinary core idea.

Another possibility is that the nature of the phenomenon itself explained why students had more trouble supplying with the Finches task. it may have been easier for students to make sense of a problem involving human-driven selective pressures, relative to environmental driven ones. The possibility is underscored by research suggesting how difficult it is for students to imagine causes that are emergent, rather than driven by agents with purposes (Hmelo-Silver, Marathe, & Liu, 2007). In the case of the Swallows tasks, students might well have made sense of the selective pressure more easily, because it conformed to the kind of causal reasoning that they may have been accustomed to using but that is different from evolutionary reasoning when environmental pressures are at play.

Last, students struggled with prediction questions, particularly on the Finches task. In the Swallows task, students had to make a fairly simple extrapolation related to traits. Inferring changes to the environment from changes to trait distribution proved much harder to do for students as was required for Finches.

Discussion

Overall, we found that three of our four tasks were sensitive to opportunities to learn students encountered in problem-based teaching. When compared to performance on pretest, nearly all test combinations showed improvements, except for one task (Finches) that proved particularly challenging for students. The tasks, importantly, tested students' mastery of performance standards that were the focus of an extended phenomenon-based unit by presenting students with a phenomenon they had not yet seen. In addition, the task performance yielded significantly

different findings across classrooms, indicating the tasks' sensitivity to differences in classroom contexts. Thus, our findings point to the possibility of designing three-dimensional assessment tasks that could be used to evaluate the success of phenomenon-based teaching in science.

At the same time, our results show that the particular task students saw mattered for their performance. Nearly all of the task pair gains were significantly different from one another. This presents a challenge for using such tasks in evaluation: it is difficult to present multiple multicomponent tasks to students because each task requires between 20-30 minutes for students to complete. Thus, while a strategy of presenting equivalent but different tasks to students in pre- and post-tests is a *potential* solution to such a problem, our study does not provide supporting evidence that this solution is viable, at least with our tasks as currently written.

Our qualitative analysis of tasks after reviewing results showed us aspects of questions that we should have anticipated would matter but did not. For example, the nature of the phenomenon as involving human- or environment-driven evolution should have led us to consider more carefully this attribute of a phenomenon as mattering for student performance, especially given prior research on student difficulties with concepts of complex causality. This discovery has led our team to adjust its design process, so that we are explicitly considering the nature of the phenomenon when developing new tasks, so that they are likely to be more comparable. In addition to considering underlying mechanisms that could differ, other considerations related to the nature of the phenomenon including whether they are macro- or micro- in space and time, whether human activity is central to the phenomenon, and the degree to which the phenomenon is culturally or societally significant (Penuel, 2018; Suárez & Bell, 2019).

Another adjustment we have made to our design process has been to develop assessments in pairs, so that we improve the degree to which each assessment's prompts are mirrored in a task

we design to be equivalent. Taking this step has facilitated new kinds of conversations at the design stage about equivalence of tasks. At the same time, it has raised new questions for us about the challenges of producing parallel tasks with different phenomena and using datasets that are not entirely parallel. It has, furthermore, required us to identify needs for additional kinds of data for some phenomena not only to meet the demands of a performance expectation but also to ensure tasks are parallel. In a task we designed related to carrying capacity in ecosystems, for example, creating parallel tasks required us to find climate data for one set of ecosystems that had not been part of the scientific study we used to develop the task, because climate is a factor in carrying capacity and because the study used for the parallel task did include such data.

With respect to equity, there continue to be opportunities to improve our assessment design process, as well as our data collection and analysis process. With respect to design, although we use a survey of student interest to select phenomena that anchor our units (Penuel, Reiser, et al., 2018), we have not done so for assessments. It is, however, possible to use phenomena identified in those interest surveys that were not chosen to anchor units as the basis for assessment tasks. In addition, the finding that students who are Latinx scored slightly lower on our assessments is concerning to us. Many Latinx students in the district are also emerging bilinguals. Because we did not collect data on students' home language, we cannot determine readily whether this difference in performance was due to cultural bias in the ways that students approached the phenomenon or interpreted questions or whether the difference was to the accessibility of tasks to emerging bilingual students. Future data collection and analysis efforts will need to provide us with more evidence regarding how each of these dimensions (cultural bias, language) could be contributing to student performance.

Another related, ongoing concern in our design deliberations is the role of scaffolding in assessments. On the one hand, scaffolding can and has been used effectively in tasks to help students from widely varying backgrounds to gain access to complex phenomena and to tasks (e.g., Kang et al., 2014). This includes linguistic scaffolding in tasks, in which descriptions of phenomena are presented in students' home language (Buxton et al., 2014). A concern, however, is that providing hints or advanced organizers for student responses can lower the cognitive demand of an assessment task. Indeed, many such scaffolds do just this (Tekkumru-Kisa & Stein, 2015). A challenge is to determine when a scaffold supports gaining access to tasks, without lowering cognitive demand. In part, this is a matter of design but also depends on empirical study of student responses to tasks with different levels and types of scaffolding.

Conclusions

Our study illustrates one approach to the design and validation of assessments of students' three-dimensional learning for purposes of program evaluation. Using an evidence-centered design process we constructed four multicomponent assessment tasks of student learning that were anchored in phenomena, required students to use three dimensions, and that could be administered in a single class period. The assessments met the criterion of being "proximal transfer" tasks in that they required students to apply knowledge of core ideas, practices, and crosscutting concepts to answer questions related to a phenomenon they had never seen before. When combined into test forms including two tasks, pre- and post-assessments yielded significant gains, showing evidence of sensitivity to instruction. Moreover, these gains were not different for students of different genders or ethnicity, showing no evidence of bias in overall performance levels.

Our findings also point to the ongoing challenges of developing three-dimensional assessments. The tasks were not equally easy or difficult, as demonstrated by our finding that task combination is a significant predictor of student performance. This finding is consistent with past research showing how important task context can be for student performance in extended tasks in science (Settlage & Jensen, 1996; Shavelson, Baxter, & Pine, 1991). There are clear implications for use of such tasks to evaluate curriculum—even when tasks are closely aligned to content taught—because the choice of assessment phenomenon or problem could lead to over- or under-estimation of the efficacy of materials.

Even with carefully chosen phenomena and templates for task design, three-dimensional assessment design can yield assessments that are of widely varying quality. Some may fail to sufficiently engage students of different ethnicities, genders, and linguistic backgrounds. Others may not be sufficiently connected to students' everyday lives in their families and communities, and thus fail to be culturally relevant. Still others may not be sufficiently demanding, given the standard being assessed. And, if demanding enough to match the expectation of the standard, students may have gaps in understanding that make it difficult for them to perform well, even after exposure to curriculum activities intended to develop their understanding. Finally, some may yield strong alignment to policy guidance for defining aspects of the three dimensions, but not necessarily with research on disciplinary core ideas, science and engineering practices, and crosscutting concepts. This will no doubt become an increasing challenge in the future, as critiques emerge of how particular practices are framed in the standards (e.g., Gunckel & Tolbert, 2018) and as studies like ours and others like it add to the knowledge base about three-dimensional learning.

Examples of valid proximal transfer tasks are important to the field, because they can help evaluate problem-based teaching with curricular resources when more distal tasks are not available. Ideally, evaluations of curriculum materials draw on a wide variety of evidence that is both distant from and close to classroom instruction (Ruiz-Primo et al., 2002). However, state tests that can evaluate efforts to promote three-dimensional learning are still under developed. In addition, even when such tests are available, they are not always available for the grades being tested or adequate to address the particular content of units. Therefore, valid proximal tasks will continue to play a critical role in the future for identifying effective strategies for promoting three-dimensional learning.

There are a number of potential implications from our study findings for assessment design, not just for evaluation purposes but also for assessments used to monitor performance of schools and districts. First, the recommendation by the National Research Council (2014) to use matrix sampling for monitoring assessments—that is, assessments that use a systematic method to assign samples of tasks to different samples of students—to ensure representativeness of the domain of standards may be critical. Though our assessments were designed for a different purpose, our extended tasks are similar in form to some of the emerging state-level tests, and so lessons learned likely apply. For even high-quality extended tasks, it is difficult to sufficiently address even a single performance expectation. It would be difficult to assess each standard taught in a given year adequately, and likely impossible to do so on the tests typically given in states that are intended to assess multiple years of instruction. Second, to ensure comparability of tasks, it may be necessary to empirically evaluate what task features tend to be more difficult than others. Empirical research to validate hypothetical learning progressions presented in *A Framework for K-12 Science Education* (National Research Council, 2012) and as elaborated in

supplemental appendices of the NGSS would be another useful endeavor with important assessment implications. Such research could also help identify the conditions under which we could claim that students are able to apply learning from one phenomenon in ways that reflects generalized understanding of disciplinary core ideas, practices, and crosscutting concepts. adequately generalize from phenomenon-based learning experiences. Finally, to ensure that assessment tasks are fair and equitable, we will need to gather more and consistent data not only on opportunities to learn as experienced in classrooms, but also gather data on students' interest and engagement in assessment tasks. It is a significant limitation of this study that our opportunity to learn data are incomplete. Opportunity to learn data are critical to establishing that students have a good chance of performing well on tests. Tasks, moreover, should be interesting to students and reflect an understanding that all learning—including science learning—is a cultural process that can be enhanced when students' everyday experiences are leveraged (National Academies of Sciences Engineering and Medicine, 2018a).

More broadly, if the Next Generation Science Standards promote three-dimensional learning and encourage the use of relevant and accessible phenomena and problems as anchors, then the ways we assess this learning must retain integrity to these goals. Thus, analyses of efforts like this one to develop validity evidence for such assessment are needed, to provide a possible pathway to further development and refinement of three- dimensional assessments.

EndNotes

[1]With few exceptions (those focused on Engineering, Technology, and Society), all standards integrate the three dimensions; therefore, for the remainder of the paper, we refer to assessments of the NGSS as "three-dimensional" or 3D assessments. *A Framework for K-12 Science*

Education (National Research Council, 2012) provides definitions for each of the dimensions used in this report. *Disciplinary Core Ideas* refer to those ideas that have broad importance across multiple science and engineering disciplines or are a key organizing principle of a discipline, provide a tool for investigating more complex phenomena and problems, relate to the interests, experiences, and concerns of students and their communities that require science knowledge, and are teachable and learnable across grades at increasing levels of sophistication (p. 31). *Science and Engineering Practices* are the principal practices scientists use to build models and theories about the natural world and that engineers use to design solutions to solve problems (p. 30). The *Crosscutting Concepts* are ideas that have application across multiple domains of science, and they provide a way to link multiple disciplinary ideas together (p. 30). We recognize that in the field, there are many different interpretations of the crosscutting concepts and their significance for instruction and assessment. We refer readers to Rivet and colleagues' (Rivet, Weiser, Lyu, Li, & Rojas-Perilla, 2016) review of some of these meanings.

[2]These are all examples of phenomena that anchor the biology curriculum for which these assessments were developed. For more on the first phenomenon and how it figures in instruction, see (National Academies of Sciences Engineering and Medicine, 2018, pp. 90-102).

References

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36-46.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do? In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Dordrecht, the Netherlands: Springer.

Achieve. (2015). *NGSS evidence statements*. Washington, DC: Author.

Achieve. (2016). *EQuIP rubric for lessons and units: Science, Version 3.0*. Washington, DC: Author.

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

Achieve. (2018a). *Current approaches to systems of assessment in science: Themes and models*. Washington, DC: Author.

Achieve. (2018b). *Science task screener*. Washington, DC: Author.

Achieve. (2019). *The state of state science education policy: Achieve's 2018 Science policy survey*. Washington, DC: Author.

Alonzo, A. C., & Ke, L. (2016). Taking stock: Existing resources for assessing a new vision of science learning. *Measurement: Interdisciplinary Research and Perspectives, 14*(4), 119-152.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Beall, C. M. (2007). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proceedings of the National Academy of Sciences, 104*(1), 8655-8660.

Beall, C. M., Song, K., Elston, R. C., & Goldstein, M. C. (2004). Higher offspring survival among Tibetan women with high oxygen saturation genotypes residing at 4,000M. *Proceedings of the National Academy of Sciences, 101*(39), 14300-14304.

Blank, R. K., & Adams, E. (2018). *How well aligned are state assessments with state standards or Common Core/Next Generation Science Standards?* Chicago, IL: NORC.

Blumenfeld, P., Soloway, E., Marx, R. W., Guzdial, M., & Palincsar, A. S. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26*(3&4), 369-398.

Boag, P. T., & Grant, P. R. (1981). Intense natural selection in a population of Darwin's finches (Geospizinae) in the Galápagos. *Science, 214*(4516), 82-85.

Brown, C. R., & Brown, M. B. (2013). Where has all the road kill gone? *Current Biology, 23*(6), 233-234.

Buxton, C. A., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S.-h., Choi, Y.-j., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix, 2*, 1-21.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, assessment, and professional development*. Mahwah, NJ: Erlbaum.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2015). Building an assessment argument to design and use NGSS assessments to evaluate the efficacy of curriculum interventions. *American Journal of Evaluation, 37*(2), 174-192.

Engle, R. A. (2012). The productive disciplinary engagement framework: Origins, key concepts, and developments. In Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 161-200). New York, NY: Routledge.

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item deatures on the Assessment of students' written scientific explanations. *Research in Science Education, 45*, 527–553.

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the Next Generation Science Standards. *Science Education, 99*(6), 1041-1048.

Francis, C., Breland, T. A., Østergaard, E., Lieblein, G., & Morse, S. (2013). Phenomenon-based learning in agroecology: A prerequisite for transdisciplinarity and responsible action. *Transdisciplinarity and Responsible Action, Agroecology and Sustainable Food Systems, 37*(1), 60-75.

Gibbs, H. L., & Grant, P. R. (1987). Oscillating selection on Darwin's finches. *Nature, 327*, 511-513.

Gunckel, K. L., & Tolbert, S. (2018). The imperative to move toward a dimension of care in engineering education. *Journal of Research in Science Teaching, 55*(7), 938-961.

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

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235-266.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences, 16*(3), 307-331.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education, 98*(4), 674-704.

Kearns, A. M., Joseph, L., White, L. C., Austin, J. J., Baker, C., Driskell, A. C., Malloy, J. F., & Omland, K. E. (2016). Norfolk Island Robins are a distinct endangered species: Ancient

DNA unlocks surprising relationships and phenotypic discordance within the Australo-Pacific Robins. *Conservation Genetics, 17*, 321-355.

Kelly, G. J. (2012). The social bases of disciplinary knowledge and practice in productive disciplinary engagement. *International Journal of Educational Research, 64*, 211-214.

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

Kolodner, J. L., Starr, M. L., Edelson, D., Hug, B., Kanter, D. E., Krajcik, J., Lancaster, J. A., Laster, T. A., Leimberer, J., & Reiser, B. J. (2008). Implementing what we know about learning in a middle-school curriculum for widespread dissemination: The Project-based Inquiry Science (PBIS) story. In P. A. Kirschner, F. Prins, V. Jonker, & G. Kanselaar (Eds.), *Proceedings of the 8th International Conference of the Learning Sciences* (Vol. 3, pp. 274-281). Utrecht, the Netherlands: Erlbaum.

Lazarus, S. S., Thurlow, M. L., Lail, K. E., & Christensen, L. (2009). A longitudinal analysis of state accommodations policies: Twelve years of change 1993-2005. *Journal of Special Education, 43*(2), 67-80.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115-136.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday classrooms and professional settings* (pp. 251-277). Mahwah, NJ: Erlbaum.

Lu, J., Bridges, S., & Hmelo-Silver, C. E. (2014). Problem-based learning. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 298-318). New York, NY: Cambridge.

Manz, E. (2015a). Representing argumentation as functionally emergent from scientific activity. *Review of Educational Research, 85*(4), 553-590.

Manz, E. (2015b). Resistance and the development of scientific practice: Designing the mangle into science instruction. *Cognition and Instruction, 33*(2), 89-124.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

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

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling. *Journal of Research in Science Teaching, 46*(6), 675-698.

National Academies of Sciences Engineering and Medicine. (2018a). *How people learn II: Learners, cultures, and contexts*. Washington, DC: National Academies Press.

National Academies of Sciences Engineering and Medicine. (2018b). *Science and engineering for grades 6-12: Investigation and design at the center*. Washington, DC: National Academies Press.

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

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

Nehm, R. H., Beggrow, E., Opfer, J., & , & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher, 74*(2), 92-98.

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

Odden, O. B., & Russ, R. S. (2019). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education, 103* (1), 187-205.

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

Pellegrino, J. W., DiBello, L., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologst, 51*(1), 59-81.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10-18.

Penuel, W. R. (2014). Studying science and engineering learning in practice. *Cultural Studies of Science Education*, 1-16. doi:10.1007/s11422-014-9632-x

Penuel, W. R. (2018). *PeACESSE Resource E: Selecting compelling anchoring phenomena for equitable science teaching*. Seattle, WA: STEM Teaching Tools, University of Washington.

Penuel, W. R., Frumin, K., Van Horne, K., & Jacobs, J. K. (2018, April). *A phenomenon-based assessment system for three-dimensional science standards: Why do we need it and what can it look like in practice?* . Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Penuel, W. R., Harris, C. J., D'Angelo, C., DeBarger, A. H., Gallagher, L. P., Kennedy, C. A., . . . Krajcik, J. S. (2015). Impact of project-based curriculum materials on student learning in science: Results of a randomized controlled trial. *Journal of Research in Science Teaching, 52*(10), 1362-1385.

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

Penuel, W. R., Reiser, B. J., Novak, M., McGill, T., Frumin, K., Van Horne, K., Sumner, T., & Watkins, D. A. (2018, April). *Using co-design to test and refine a model for three-dimensional science curriculum that connects to students' interests and experiences*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (pp. 787-851). Washington, DC: AERA.

Petrosino, A. J. (1998). *The use of reflection and revision in hands-on experimental activities by at risk children.* Unpublished dissertation. Vanderbilt University, Nashville, TN.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13*(3), 273-304.

Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263-305). Mahwah, NJ: Lawrence Erlbaum.

Rivet, A. E., Weiser, G., Lyu, X., Li, Y., & Rojas-Perilla, D. (2016). What are crosscutting concepts in science? Four metaphorical perspectives. 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. 970-973). Singapore: International Society of the Learning Sciences.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal Design for Learning*. Washington, DC: ASCD.

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., Passmore, C., & Reiser, B. J. (2017). Moving beyond 'knowing about' science to making sense of the world. In C. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 3-21). Washington, DC: NSTA.

Settlage, J., & Jensen, M. (1996). Investigating the inconsistencies in college student responses to natural selection test questions. *Electronic Journal of Science Education, 1*(1).

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-362.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32-37.

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.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers assessment literacy. *Journal of Science Teacher Education, 22*, 371-391.

Suárez, E., & Bell, P. (2019, April). *Supporting expansive science learning through different classes of phenomena*. Paper presented at the NARST Annual International Conference, Baltimore, MD.

Symeonidis, V., & Schwarz, J. F. (2016). Phenomenon-based teaching and learning through the pedagogical lenses of phenomenology: The recent curriculum reform in Finland. *Forum Oświatowe, 28*(1), 31-47.

Tan, E., & Calabrese Barton, A. (2012). *Empowering science and mathematics education in urban schools*. Chicago, IL: University of Chicago Press.

Tekkumru-Kisa, M., & Stein, M. K. (2015). Learning to see teaching in new ways: A foundation for maintaining cognitive demand. *American Educational Research Journal, 51*(1), 105-136.

Thurlow, M. L. (2012). Students with disabilities, testing accommodations. In J. A. Banks (Ed.), *Encyclopedia of diversity in education* (pp. 2090-2092). Thousand Oaks, CA: SAGE.

Thurlow, M. L., & Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education, 39*(1), 331-369.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments*. Minneapolis, MN: National Accessible Reading Assessment Projects.

Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting assessment tasks to support three-dimensional learning. *Journal of Chemical Education, 95*, 207-217.

Walker, A., & Leary, H. (2009). A Problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *Interdisciplinary Journal of Problem Based Learning, 3*(1), 12-43.

Wertheim, J. (2016). Commentary: Taking stock: Implications of a new vision of science learning for state science assessment. *Measurement: Interdisciplinary Research & Perspective, 14*(4), 158-161.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. Palo Alto, CA: Stanford Center for Assessment, Learning, and Equity, Stanford University.

Table 1.

Student Sample Characteristics

	American Indian/ Alaskan Native	Asian	Black (Not Hispanic)	Hispanic	White (Not Hispanic)	Native Hawaiian/ Pacific Islander	Multiple Races
Male	1 (0.2%)	4 (0.7%)	13 (2.2%)	94 (16.1%)	119 (20.4%)	0 (0%)	24 (4.1%)
Female	2 (0.3%)	18 (3.1%)	31 (5.3%)	115 (19.7%)	144 (24.7%)	1 (0.2%)	17 (2.9%)

N = 583

Table 2.

Implementation Data (Student Survey)

	A (n = 8)	B (n = 139)	C (n = 17)	D (n = 194)	E (n = 266)	J (n = 104)
Coherence						
% who say the class made progress on student-generated questions related to the anchoring phenomenon	100%	100%	76%	73%	67%	78%
% who said they knew why the class did they activity they did	88%	82%	88%	84%	85%	88%
% who have ideas about what questions to address next	88%	45%	41%	60%	55%	66%
Contribution						
% who said they shared their ideas in both small group and whole class discussion	25%	17%	0%	13%	9%	26%
% who said they shared their ideas with no one	25%	11%	24%	3%	23%	9%
Relevance						
% who said the day's lesson was relevant to them personally	63%	57%	41%	49%	40%	59%
% who said the day's lesson was not relevant to anyone at all	13%	4%	6%	5%	9%	9%

Table 3.

Overview of Tasks Presented to Students

Task Name and Phenomenon	**Focal Performance Expectation(s)**	**Components of DCIs, Science and Engineering Practices, and Crosscutting Concepts Assessed**
Evolution of Swallows Microevolution of wing length in population of swallows that adapted to life above a busy highway	**HS-LS4-3** Apply concepts of statistics and probability to support explanations that organisms with an advantageous heritable trait tend to increase in proportion to organisms lacking this trait **HS-LS4-4** Construct an explanation based on evidence for how natural selection leads to adaptation of populations.	*Claims Related to DCI Used to Select Phenomenon (from* Framework): The traits that positively affect survival are more likely to be reproduced and thus are more common in the population. Natural selection leads to adaptation—that is, to a population dominated by organisms that are anatomically, behaviorally, and physiologically well suited to survive and reproduce in a specific environment. Natural selection leads to adaptation, that is, to a population dominated by organisms that are anatomically, behaviorally, and physiologically well suited to survive and reproduce in a specific environment. That is, the differential survival and reproduction of organisms in a population that have an advantageous heritable trait leads to an increase in the proportion of individuals in future generations that have the trait and to a decrease in the proportion of individuals that do not. *Analyzing and Interpreting Data:* Apply concepts of statistics and probability (including determining function fits to data, slope, intercept, and correlation coefficient for linear fits) to scientific and engineering questions and problems, using digital tools when feasible *Constructing Explanations:* Construct an explanation based on valid and reliable evidence obtained from a variety of sources (including

		~~students' own investigations, models,~~ theories~~, simulations,~~ peer review) and the assumption that theories and laws that describe the natural world operate today as they did in the past and will continue to do so in the future. *Patterns:* Different patterns may be observed at each of the scales at which a system is studied and can provide evidence for causality in explanations of phenomena *Cause and Effect:* Empirical evidence is required to ~~differentiate between cause and correlation and~~ make claims about specific causes and effects.
Galápagos Ground Finches Microevolution of beak and wing length among finches on the Galápagos Islands that adapted to reduced food availability	**HS-LS4-3** Apply concepts of statistics and probability to support explanations that organisms with an advantageous heritable trait tend to increase in proportion to organisms lacking this trait **HS-LS4-4** Construct an explanation based on evidence for how natural selection leads to adaptation of populations.	Same as for Evolution of Swallows
Two Species or One? Resolving uncertainty regarding whether a	**HS-LS4-1:** Communicate scientific information that common	*Claims Related to DCI Used to Select Phenomenon (from* Framework):

population of birds us part of an endangered species on the basis of genetic similarity	ancestry and biological evolution are supported by multiple lines of evidence **HS-LS4-5:** Evaluate the evidence supporting claims that changes in environmental conditions may result in (1) increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species [cause and effect]	Genetic information, like the fossil record, provides evidence of evolution. DNA sequences vary among species, but there are many overlaps; in fact, the ongoing branching that produces multiple lines of descent can be inferred by comparing the DNA sequences of different organisms. Changes in the physical environment, whether naturally occurring or human induced, have thus contributed to the expansion of some species, the emergence of new distinct species as populations diverge under different conditions, and the decline — and sometimes the extinction — of some species Species become extinct because they can no longer survive and reproduce in their altered environment. *Obtaining, Evaluating, and Communicating Information:* Communicate scientific phenomenon in multiple formats (including orally, graphically, textually, and mathematically) *Engaging in Argument from Evidence:* Evaluate the evidence behind currently accepted explanations o solutions to determine the merits of arguments *Patterns:* Different patterns may be observed at each of the scales at which a system is studied and can provide evidence for causality in explanations of phenomena *Cause and Effect:* Empirical evidence is required to ~~differentiate between cause and correlation and~~ make claims about specific causes and effects
Human Adaptation on the Tibetan Plateau	~~**HS-LS4-3:**~~	*Claims Related to DCI Used to Select Phenomenon (from* Framework):

Adaptation of human beings to life at high altitudes evident in size of blood vessels	~~Apply concepts of statistics and probability to support explanations that organisms with an advantageous heritable trait tend to increase in proportion to organisms lacking this trait [patterns]~~ **HS-LS-4-4:** Construct an explanation based on evidence for how natural selection leads to adaptation of populations [Cause and effect]	Natural selection leads to adaptation, that is, to a population dominated by organisms that are anatomically, behaviorally, and physiologically well suited to survive and reproduce in a specific environment. That is, the differential survival and reproduction of organisms in a population that have an advantageous heritable trait leads to an increase in the proportion of individuals in future generations that have the trait and to a decrease in the proportion of individuals that do not. *Constructing Explanations:* Construct an explanation based on evidence obtained from a variety of sources (including ~~students' own investigations, models,~~ theories~~, simulations,~~ peer review) and the assumption that theories and laws that describe the natural world operate today as they did in the past and will continue to do so in the future *Cause and Effect:* Empirical evidence is required to ~~differentiate between cause and correlation and~~ make claims about specific causes and effects.

*Crossed out text indicates text from the foundations boxes of the Next Generation Science Standards that we chose not to assess as part of the task.

Table 4.

Test Form Construction

	Form A		**Form B**		**Form C**	
Task	Version 1	Version 2	Version 1	Version 2	Version 1	Version 2
Evolution of Swallows		X (1st)	X (2nd)			X (2nd)
Galápagos Ground Finches	X (2nd)			X (1st)		X (1st)
Two Species or One?	X (1st)		X (1st)		X (1st)	
Human Adaptation on the Tibetan Plateau		X (2nd)		X (2nd)	X (2nd)	

Table 5.

Maximum Points by Task and by Test Form

Test Form	Task 1	Task 2	Maximum Score for Task 1	Maximum Score for Task 2	TOTAL Possible Score
A1	Species	Finches	13	13	26
A2	Swallows	Tibet	14	14	28
B1	Species	Swallows	13	14	27
B2	Finches	Tibet	13	14	27
C1	Species	Tibet	13	14	27
C2	Finches	Swallows	13	13	27

Table 6.

Scoring Guide for *Human Adaptation on the Tibetan Plateau*, Question 1

Question 1.

Using the information in the table about people visiting the mountains and Tibetans living at high elevations, construct an explanation for why, over many generations, the people of Tibet today are physiologically different from people living at lower elevations.

POSSIBLE POINTS = 4

Points	**Required Components of Answer**	**Examples**
+1	Accurately describes a trait in the population (The population can either refer to the Tibetans or the Lowlanders visiting.)	The heart rate and blood vessels are different from lowlanders.
+1	Connects a specific advantageous trait to a better chance of survival at high altitudes or says those with specific less advantageous traits might not survive as well.	A response we see today that could have provided an advantage was the number of blood cells produced. They would be able to get more oxygen to their cells and have better stamina than the others which would be an advantage for survival. Tibetans adapted to have lower heart rate. Those who adapted were able to survive in this elevation.
+1	Connects surviving to reproducing or talks about passing on advantageous traits. Have to make an explicit reference to survival (along with reproduction/passing on traits).	Over many generations, the people of Tibet probably adapted to the environment & passed on the traits for survival.
+1	Says that the difference between highlanders and lowlanders is caused by natural selection/adaptation (must accurately use one of these vocab words).	Their ancestors' bodies adapted to the high elevation and passed that down to their children.

Table 7.

Pre and Post-Test Scores for Each Assessment Task

Task Order	Task	Mean	Std. Dev.	Min	25th	Median	75th	Max
Pre	Finch1	0.23	0.20	0.00	0.10	0.20	0.40	0.90
	Finch2	0.19	0.20	0.00	0.08	0.10	0.30	0.70
	Swallows1	0.34	0.22	0.00	0.10	0.30	0.50	0.90
	Swallows2	0.28	0.21	0.00	0.10	0.20	0.40	0.90
	Species1	0.29	0.16	0.00	0.20	0.30	0.40	0.60
	Tibet2	0.34	0.23	0.00	0.15	0.30	0.50	1.00
Post	Finch1	0.34	0.23	0.00	0.10	0.40	0.50	0.90
	Finch2	0.30	0.22	0.00	0.10	0.30	0.50	0.80
	Swallows1	0.42	0.22	0.00	0.20	0.40	0.60	0.90
	Swallows2	0.40	0.24	0.00	0.20	0.40	0.60	0.90
	Species1	0.39	0.18	0.00	0.30	0.40	0.50	0.90
	Tibet2	0.47	0.22	0.00	0.30	0.50	0.65	1.00

Table 8.

Gain Scores for Each Teacher

Teacher	Mean	Std. Dev.	Min	25th	Median	75th	Max
A	0.08	0.13	-0.10	-0.02	0.05	0.18	0.35
B	-0.08	0.17	-0.55	-0.15	-0.05	0.00	0.20
C	0.03	0.08	-0.10	0.00	0.05	0.05	0.15
D	-0.06	0.13	-0.40	-0.14	-0.10	0.05	0.20
E	0.00	0.13	-0.10	-0.07	-0.05	0.05	0.15
F	0.14	0.13	0.00	0.05	0.10	0.25	0.35
G	0.02	0.12	-0.20	-0.02	0.00	0.06	0.25
H	0.04	0.16	-0.60	-0.05	0.05	0.15	0.30
I	-0.05	0.17	-0.40	-0.14	0.00	0.08	0.15

Table 9.

Numerical Comparison of Pre and Post Tasks by Gain Score

Task Combination (Pre then Post)	N	Mean (%)	SD (%)
Tibet2 Finch2	91	-5	25
Swallows1 Finch2	92	-4	20
Species1 Finch1	98	2	19
Swallows1 Species1	94	4	21
Tibet2 Finch1	91	4	28
Tibet2 Species1	182	4	23
Tibet2 Swallows2	180	7	27
Species1 Swallows1	90	14	23
Finch1 Swallows2	92	16	25
Finch1 Species1	93	17	22
Swallows2 Tibet2	185	18	26
Species1 Tibet2	187	18	24
Finch1 Tibet2	87	21	26

Finch2 Swallows1	93	23	24
Finch2 Tibet2	91	29	25

Table 10.

Regression Results

		Regression Results		
		Task	Task/Teacher	Task/Teacher/ Gender/Ethnicity
Task Combination	Intercept (Tibet2_Finch2, A, Male, White)	-0.044*	-0.004	0.021
		(0.026)	(0.031)	(0.033)
	Swallows1_Finch2	0.004	0.004	0.004
		(0.036)	(0.035)	(0.035)
	Species1_Finch1	0.061*	0.069**	0.066*
		(0.035)	(0.034)	(0.034)
	Swallows1_Species1	0.078**	0.078**	0.078**
		(0.036)	(0.034)	(0.034)
	Tibet2_Finch1	0.079**	0.090***	0.089**
		(0.036)	(0.035)	(0.035)
	Tibet2_Species1	0.084***	0.089***	0.087***
		(0.031)	(0.030)	(0.030)
	Tibet2_Swallows2	0.106***	0.115***	0.114***
		(0.031)	(0.030)	(0.030)
	Species1_Swallows1	0.186***	0.187***	0.185***
		(0.036)	(0.035)	(0.035)
	Finch1_Swallows2	0.204***	0.210***	0.209***
		(0.036)	(0.034)	(0.034)
	Finch1_Species1	0.216***	0.222***	0.220***
		(0.036)	(0.034)	(0.034)
	Swallows2_Tibet2	0.226***	0.231***	0.228***
		(0.031)	(0.030)	(0.030)
	Species1_Tibet2	0.227***	0.232***	0.229***
		(0.031)	(0.030)	(0.030)

	Finch1_Tibet2	0.252***	0.253***	0.251***
		(0.037)	(0.035)	(0.035)
	Finch2_Swallows1	0.276***	0.276***	0.275***
		(0.036)	(0.034)	(0.034)
	Finch2_Tibet2	0.331***	0.332***	0.330***
		(0.036)	(0.035)	(0.035)
Teacher	B		-0.146***	-0.152***
			(0.025)	(0.025)
	C		0.006	0.022
			(0.031)	(0.033)
	D		-0.071***	-0.053**
			(0.024)	(0.027)
	E		-0.018	-0.016
			(0.040)	(0.040)
	F		0.128***	0.129***
			(0.027)	(0.029)
	G		0.050	0.045
			(0.032)	(0.032)
	H		-0.019	-0.026
			(0.028)	(0.028)
	I		-0.073***	-0.079***
			(0.025)	(0.025)
	J		-0.099***	-0.103***
			(0.029)	(0.029)
	K		-0.061**	-0.068***
			(0.025)	(0.025)
Gender	Female			-0.014
				(0.011)
Ethnicity	American Indian/ Alaskan Native			0.090
				(0.077)

Asian			-0.005
			(0.031)
Black (Not Hispanic)			-0.017
			(0.021)
Hispanic			-0.034**
			(0.017)
Native Hawaiian/			-0.256
Pacific Islander			(0.165)
Multiple Races			-0.023
			(0.025)
N	1,734	1,734	1,734
Adjusted R^2	0.123	0.197	0.198
Residual Std. Error	0.242 (df = 1719)	0.231 (df = 1709)	0.231 (df = 1702)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

The differential survival and reproduction of organisms in a population that have an advantageous heritable trait leads to an increase in the proportion of individuals in future generations that have the trait and to a decrease in the proportion of individuals that do not.
There are connections here to natural selection, focusing on the changes to proportions of individuals over time with organisms with different traits.

Focus is on proportion: Tie to crosscutting concept: Rate, proportion, and scale.

Emphasis is both on the increase in proportion of individuals with traits that are advantageous and decreased in organisms without the trait. Here, over time is operationalized as being over successive generations.

Adaptation also means that the distribution of traits in a population can change when conditions change.
Conditions changing are one thing that can change what is selected for.

A change in what traits are selected for results in different adaptations.

That change also shifts the relative distribution of traits in a population.

*Bold text comes from *A Framework for K-12 Science Education* (NRC, 2012)
Figure 1.

Sample Claims and Sub-claims from analysis of PEs for LS4C

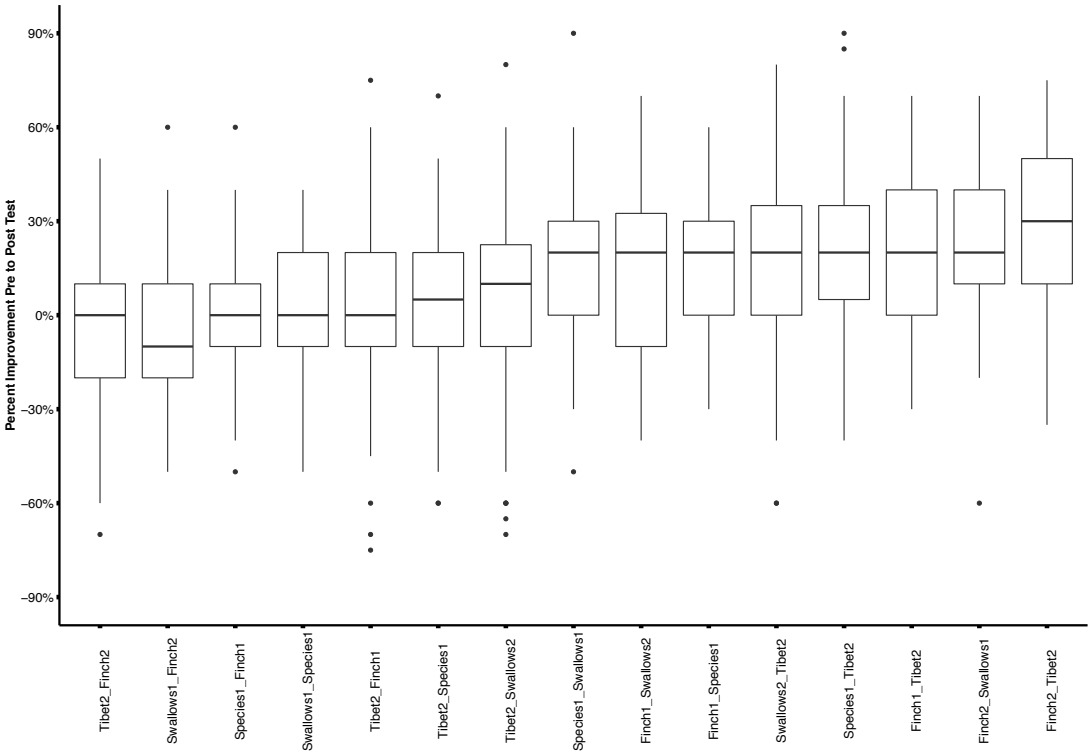

Figure 2. Boxplot of Score Growth from Pre to Post for Each Task Combination