Bridging the Gap: Evaluating a Design Approach for Curriculum-Neutral NGSS Benchmark Assessments in Middle School

Maia K. Binding and Lauren Brodsky

Abstract

This study focuses on a project designed to develop a comprehensive suite of curriculum-neutral benchmark assessments for the middle school NGSS performance expectations. Two key questions are addressed, "*Does the project approach lead to the design of items that elicit three-dimensional student performances aligned to the NGSS PEs? Are these items curriculum-neutral in that they are accessible to students regardless of their curriculum of instruction?*" Findings from four evidence sources are analyzed and discussed: 1) multi-curriculum review, 2) cognitive lab interviews, 3) external expert review, 4) classroom field tests. Results indicate that the project design and development approach results in assessment items that are curriculum neutral and elicit three-dimensional responses in line with the NGSS PEs. The resulting items are proposed for use in a broad range of classrooms, regardless of curriculum choice, to assess authentic student understanding of the three dimensions of the PEs.

Background and Problem

The advent of the Next Generation Science Standards (NGSS Lead States, 2013) and the Framework for K-12 Science Education (NRC, 2012) has heralded the opportunity for a promising, significant shift in science teaching and learning in the United States. The NGSS move from teacher-centered, content knowledge acquisition to students developing and applying knowledge in new situations (Krajcik et al., 2014). The NGSS address a smaller number of topics than previous standards, but with more depth (NGSS Lead States, 2013), and present science as a process rather than a product (NRC, 2012). Intertwined with this depth of coverage are three dimensions of content: science and engineering practices (SEPs), crosscutting concepts (CCCs), and disciplinary core ideas (DCIs). This shift in thinking requires a parallel change in how students' understanding of these three dimensions is assessed.

A complete assessment system should include a range of assessments at all levels, from classroom assessments embedded in specific curricula, to classroom benchmark assessments, to state-level tests typically administered every few years (NRC, 2014; Osborne et al 2016; Shepard et al, 2017). As the need for NGSS-designed assessments has become clear, several projects have developed exemplars for three-dimensional assessments, while others have developed three-dimensional assessments specific to their NGSS-designed curriculum (Harris et al., 2018, MacPherson et al., 2017, SNAP, 2018, Zaidi, et al., 2018). However, schools and districts across the country are utilizing a wide range of curricula and materials. Additionally, standardized assessments are being developed at the district and state level that are not necessarily aligned to a particular curriculum. Given these circumstances, students need to be provided the opportunity to apply their understanding to assessments that are not embedded in or based on their curriculum before they take standardized assessments. The NGSS leaves room for the use of a variety of visual representations, phenomena, and aspects of the presentation of the three dimensions within curricula. Students will benefit from practice with assessments that require them to transfer their understanding to contexts that are different from what they have experienced with their particular curriculum. To this end, the Lawrence Hall of Science (The Lawrence) launched a four-year project, funded by the Carnegie Corporation of New York, to develop summative, curriculum-neutral, NGSS-designed, three-dimensional assessments for the middle school NGSS Performance Expectations (PEs) in Earth, Life, and Physical Sciences and Engineering. These model assessments are designed to serve as a bridge between curriculum and standardized assessments. This paper aims to outline the design processes, and initial review and field-testing results of this project, sharing the lessons learned thus far. In this way we will address the questions: *Does the project approach lead to the design of items that elicit three-dimensional student performances aligned to the NGSS PEs? Are these items curriculum-neutral in that they are accessible to students regardless of their curriculum of instruction?*

Assessment Item and Study Design

This work builds on work from a collaboration between The Lawrence, the American Museum of Natural History, and the University of Connecticut, funded by the National Science Foundation (NSF). This project developed a model NGSS-aligned curriculum unit, with embedded and benchmark assessments (*Disruptions in Ecosystems,* DRL 1418235). Several of the end-of-chapter assessments were identified as model assessments by the Achieve *Task Annotation Project in Science* (personal communication, Achieve, 2019). A supplement to the NSF grant provided the project team with the opportunity to score and analyze student responses to these assessments. This work provided significant insight in determining which assessment features and characteristics best elicited three-dimensional responses and should be carried into the current project.

The current project has also reviewed published model items and research on specific content areas and practices, when available (Badrinarayan et al., 2019; McElhaney et al., 2019).The assessment development approach for this project has been further informed by the work of others in the field, including the practice of evidence-centered design (Almond, et al., 2002; Mislevy & Haertel, 2006) and research and design tools and approaches for three-dimensional assessment development. These design tools and approaches include the Next Generation Science Assessment (NGSA) approach described in Harris et al., 2019, and *STEM Teaching Tool #29* (Penuel et al, 2016).

The project team consists of curriculum and assessment developers, the majority of whom are former classroom teachers. For each PE, two team members began the item development process with an unpack of the PE and development of initial learning performances which were then reviewed by the project team. The unpack and learning performances were then revised as needed, and used to develop design patterns and initial items. These were again reviewed by the full team. Once revised item drafts were complete, sample student responses and scoring guides were developed through a similar process.

In total, sets of assessment items have been developed across middle school content areas for 30 of the NGSS PEs and are in process for the remaining 29 PEs. A typical item set for one PE includes a range of 2-4 items specifically related to the learning performances developed from the unpacking of each PE. Items begin with scenarios or phenomena for students to respond to. Most of the item sets include simple diagrams or relevant artwork to enhance accessibility. Item sets cover all three dimensions of the PE being assessed.

Study Design

To address our research questions we collected data from four different sources: multi-curriculum group review, external expert review, cognitive lab interviews, and field tests. The data provides evidence to address both aspects of the study: a) do the items elicit three-dimensional responses aligned to the PEs, and b) are the assessment items accessible without prior experience with a particular curriculum. Initial evidence of validity of the items is being collected and evaluated following the framework of Pellegrino, DiBello, & Goldman (2016). The project is focusing on three of the five kinds of data described by Pellegrino et al.: expert review, student cognitive lab interviews, and limited studies of item and test performances on a sample of the items. The assessments will be available as open resources, for teachers to use and adapt to meet their specific needs. As teachers select and modify the assessments, most validity claims, especially any claims about test performance and certain aspects of inferential validity, will no longer apply. Given these conditions, the project has focused on a limited evaluation of the validity of the assessments to support principled adaptation of the assessments by educators for their student populations.

Multi-curriculum Group Review The full project team includes representatives from three established curriculum programs: Full Option Science Systems (FOSS), the Learning Design Group, and the Science Education for Public Understanding Program (SEPUP), which is also the project lead. Each step of the item design process (unpack, learning performances, design patterns, items, sample student responses and scoring guides) received a general review from all team members for PE alignment and item clarity. Individuals from each program specifically reviewed for the alignment of each step of the item design process with their curriculum. Knowing that the different curricular programs had varying interpretations of the PEs, reviewers documented instances where this occurred in the unpack of the PE. They also noted aspects of design patterns and items where significantly different graphic representations or vocabulary were used.

External Expert Review PE item sets were also sent for review by an external expert review team, then underwent a final round of review and revisions by the project team. The expert review team rated items against a rubric that included categories both for ability to elicit evidence about the three dimensions of the PE (*To what extent does the item prompt students to provide a response that integrates the three dimensions sought by the PE?*), and for general accessibility of the items (*Is the language of the prompt's phenomenon/scenario and specific question clear and appropriate for middle school students?*). Scores were given on a scale of 1-5 and comments were provided for each item within a set and also for the overall item set. Item sets for 30 PEs have been reviewed by the external expert review team, with the remaining item sets slated for review as they are developed.

Cognitive Lab Interviews Cognitive lab interviews were performed on 15 of the item sets, with more than one cognitive lab interview performed on a majority of those tested.[1] Students were recruited from a variety of classrooms using NGSS-aligned curricula, many not

[1] Cognitive labs and field testing of item sets were ongoing during Spring 2020 when the onset of the COVID-19 pandemic necessitated that all testing be halted as it required in-person interaction between either the interviewee and student or teacher and student. As of the writing of this proposal item sets for 21 PEs have been partially or fully tested. Item set development, as well as multi-curricular team and expert group review have continued.

using any of the curricula represented in the design process. Additional students were recruited through informal learning spaces (e.g. science summer camps, after school programs). Students ranged from "incoming" sixth grade students (in summer programs) to "outgoing" eighth grade students. Students were asked basic questions about the topics they had studied to determine general content familiarity and allow the interviewer to select an item set with content at least somewhat familiar to the student. The cognitive lab protocol focused on students interpreting and responding to the items as written, with follow up probes to elicit problem areas where items were not clear or where prompts were not eliciting students' full understanding. Notes were recorded about any prompts where the targeted dimensions were not elicited, and about any aspect of the item (wording, image, etc.) that lacked clarity or where students weren't able to engage with or demonstrate their PE-relevant understanding due to variations in non-relevant background knowledge.

Field tests Assessments were field tested at a variety of school sites, and with a range of students, including English language learners, educationally disadvantaged youth, students with special needs, etc. Participating teachers were recruited via numerous networks, including broad base listserves (e.g. NSTA, etc) and other similar platforms in addition to existing networks using the curricula of the groups represented by the project team. Participating teachers taught from a variety of NGSS-aligned curricula, including many curricular programs not represented by the project team. Testing occurred in classroom settings, and item sets were administered to up to150 students per PE. In addition to administering the item sets, teachers were asked to review and comment on the item sets in terms of clarity, accessibility, and alignment to PEs and their instruction. Teacher reviews were incorporated into the review process alongside expert and project team reviews. Item sets have been field tested in classrooms for 10 PEs.[1] Student responses were scored using the item-based rubrics to assess for accuracy and completeness of three-dimensional responses. In instances where PEs were field tested in multiple classrooms, student responses were pooled and treated as one sample set. Item-based rubrics scored on a scale of 0 (no response) to 4 (complete and correct response). Rubrics incorporated the three dimensions for the appropriate PE and allowed for a single score to be given per item.

Analyses and Findings

Does the project approach lead to the design of items that elicit three-dimensional student performances aligned to the NGSS PEs? The extent to which the design approach led to the design of items that elicit three-dimensional student performances aligned to the NGSS PEs was analysed by looking across the external expert review and the field test data, with additional evidence coming from the cognitive lab interviews.

The analysis by the external expert review team thus far has indicated most item sets elicit three-dimensional performances aligned with the NGSS PEs. Of the 30 item sets reviewed thus far, more than 80% of item sets rated 3 or above for the category of eliciting three-dimensional responses, with a score of 3 indicating modest revisions were suggested to improve three-dimensional alignment of the items. Field trial data indicated a range of student scores for each field tested item, where a score of 3 or 4 was indicative of a response that was partial or complete in all three dimensions. Data was analyzed for trends in scores, which might indicate need for revisions (e.g. if all students scored a "2" or below for one item, but a "3" or above for all other items in the set, or if one student scored a "3" on all but one item on which they scored a "1"). Additional expert feedback was provided through scaled responses (with elaborative comments) to a set of standard questions addressing the categories of clarity, equity/accessibility, appropriateness of scenario, and three-dimensionality for individual questions as well as alignment to the NGSS and promotion of sensemaking for the overall item set. Where specific concerns were raised in expert reviews, student responses as well as cognitive lab interview responses were analyzed where available to determine validity of said concern (e.g. expert commented that students might not understand a critical aspect of a diagram but student responses indicated no issues). Suggested revisions to improve alignment of individual items were addressed based on the review suggestions in concert with results from cognitive lab interviews and student responses from field testing. Any items that were not considered aligned were either significantly revised or replaced.

Early work on item design and accompanying cognitive lab interviews provided information on how three-dimensional performances can be elicited across content areas with the same SEP. Individual item drafts from a range of PEs were analyzed by the development team in order to isolate characteristic task features based on the SEPs to determine which features were key for engaging students with the practice along with the DCI and CCC. The development team documented common features as well as features where variation was necessary and/or appropriate across PEs, appropriate scaffolds for the practice, and other helpful design tips. These characteristic task features provided the developers with a scaffold that was grounded in the SEP and provided key structures for incorporating the relevant CCC and DCI. An example of the characteristic task features for the practice of modeling is shown in Figure 1 below.

Figure 1: Characteristic Task Features - Developing and Using Models

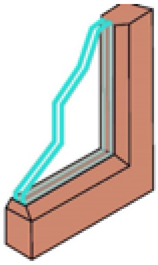
Provide a scenario/phenomenon	● 2-3 sentences ● Describe a phenomenon or set the stage for phenomenon to occur/be predicted
Provide a basic structure for the model	● Background drawing (e.g. mountain scene with lake and rivers for question around Earth Science model of water cycle), OR ● Model components (e.g. a key with simple figures and symbols to use in drawing their model), OR ● A model for a related phenomenon
Provide language that prompts model construction	● Add (arrows and) labels to the diagram below to show why/how/what happens… OR ● Add (arrows and) captions to the diagram below to explain why/how/what happens…
Prompt to use the model to explain	● Use your/the model to construct an explanation…
Possible scaffold	● Be sure to include...in your explanation/in your model

Many of these features were determined based on early student responses in cognitive lab interviews to more open modeling questions with fewer scaffolds or less specified prompting. Items that gave students data or information and asked them to “develop a model” often resulted in confusing models that were difficult to interpret and did not elicit responses from students that indicated their full level of understanding. Providing a basic structure for the model with explicit prompts and instructions to include written descriptions in the form of captions, labels, or a more lengthy explanation led students to more clearly demonstrate their three-dimensional knowledge.

A sample item taken from the item set for MS-PS 4-2 is shown in the figures below. Figure 2.a shows the item after initial development. Figure 2.b shows a sample student response from field testing of the item. Figure 2.c shows the item after revisions based on lessons learned from the development of the characteristic task features as well as from field testing and expert review.

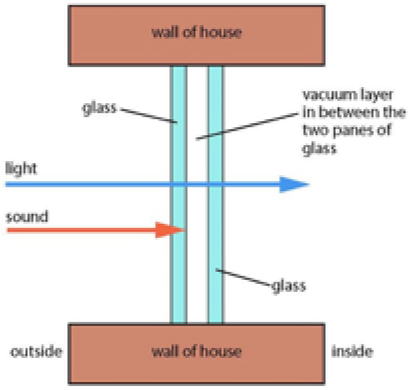
Figure 2.a: Assessment Item for MS-PS 4-2 - Initial Item Draft

Eric was at his friend’s apartment and noticed that he couldn’t hear much street noise through the window. His friend’s mother explained that the window was designed with two panes of glass separated by a vacuum layer where all the air had been removed. She drew a sketch of the window on a piece of paper.



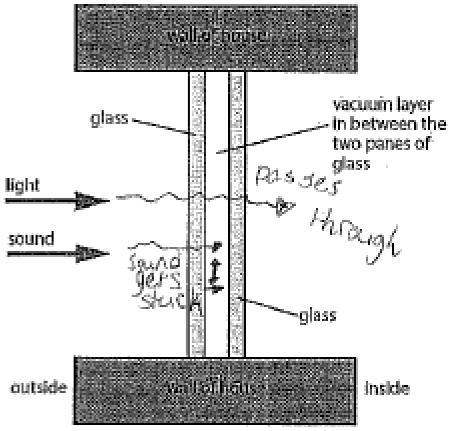
3-D sketch of the window with two panes of glass

On the diagram below, draw what happens after light and sound enter the window. Describe how the structure of the window helps to reduce the noise from the street.



Cross-section of the window in the apartment

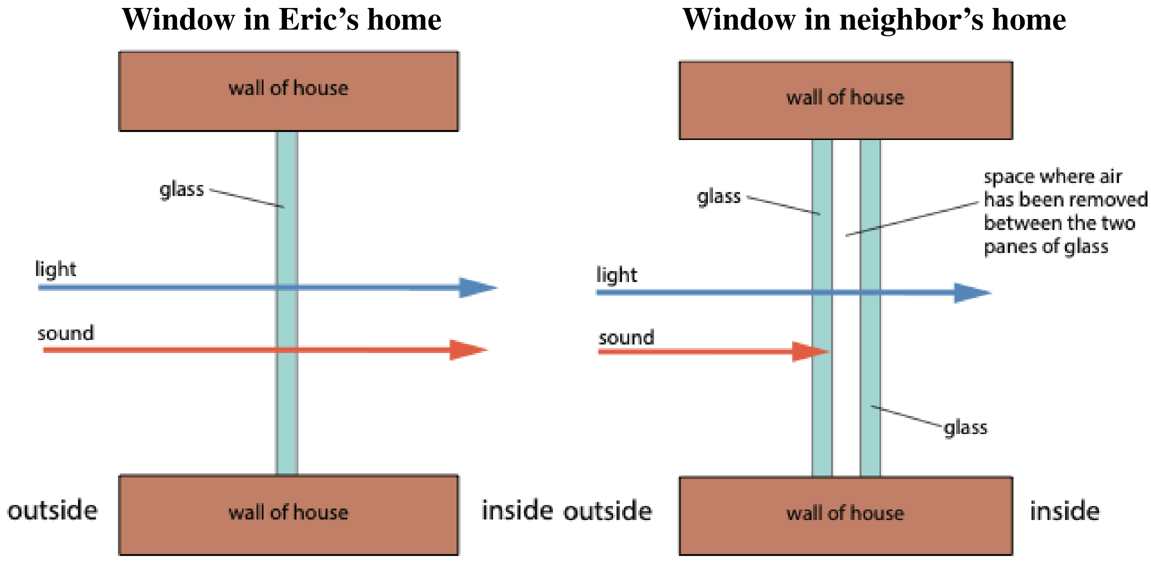
Figure 2.b: Sample Student Field Test Response for MS-PS 4-2 Item



The structure of the window helps to reduce sound from the street because the vacuum layer traps sound. Sound cannot travel through a vacuum because it is a mechanical wave, which needs matter to travel.

Figure 2.c: Assessment Item for MS-PS 4-2 - Post-Revisions

Eric was visiting his neighbor's home. He noticed that less outside noise could be heard than in his home. He compared the designs of the windows in the two homes. The ones in his home had a single pane of glass. The ones in the neighbor's home were designed with two panes of glass separated by a space where all the air had been removed.


Window in Eric's home **Window in neighbor's home**

Cross-sections of the windows in the two houses

a) On the diagrams above, draw what happens when light and sound from the outside reach the window of each house.

b) Explain why the structure of the neighbor's window helps to reduce the noise from outside more than Eric's window. Be sure to use your knowledge of sound and light as waves in your answer.

This is one of four items in the item set that assesses MS-PS 4-2, “Develop and use a model to describe that waves are reflected, absorbed, or transmitted through various materials.” In this item, as found in the development of the characteristic task features, providing a basic drawing on which students add relevant information (arrows, captions, etc) to indicate their understanding of the concepts elicits from students information related directly to the three dimensions being addressed in the PE. Students are not spending time determining how best to draw a single and double-paned window, but are instead indicating what happens to light and sound in this scenario.

Expert review of the above item gave the item a ‘4’ (minor concerns/need for revision) for all categories but appropriateness of scenario, which scored a ‘3’ (modest revisions suggested) for reasons stated above. Specific comments suggested simplification of the scenario and removal of the first sketch provided, to more tightly link the scenario to the main body of the item. These suggestions aligned with the need to better elicit the crosscutting concept of structure and function as determined from field test responses, by focusing on the comparison of single to double-paned windows. All comments were addressed in the revision process in this instance.

Additionally, in the initial stages of development, item sets occasionally treated some dimensions of the PE separately. For example, an item set might include one question that prompted the student to answer in terms of the crosscutting concept as it applied to the DCI, while another incorporated the SEP. However this led to disjointed responses in cognitive lab interviews and field test responses. Thus, the design process was adjusted to attend to all three dimensions within single items, if necessary with multiple parts, connected with a single scenario. In the examples above from the MS-PS 4-2 item set, the question was revised in multiple ways to better incorporate the crosscutting concept of structure and function and prompt for three dimensional responses. The comparison of a single and double-paned window was added to the model to elicit student responses specific to the window’s structure (one vs two panes), and the item parts were moved below the model and separated into part a) and part b) to better prompt students to respond to all parts of the question, eliciting a more three dimensional response.

The item revisions described above typify moderate to significant revisions for individual items and item sets developed thus far. A majority of items required only minor revisions similar to the separate aspects of the revisions described here in terms of clarity of language, overall prompt, and accompanying diagrams/artwork.

Are these items curriculum-neutral in that they are accessible to students regardless of their curriculum of instruction? The extent to which items are curriculum neutral was determined using curriculum group reviews that analyzed individual items and item sets to determine which would be accessible for students using a specific curriculum. The accessibility of the items for a broad audience was analyzed across the results of the external expert review, cognitive lab interviews, and field test responses.

The analysis of item sets by different curriculum groups determined that all item sets will work for a range of NGSS curricula, though occasionally individual items within a set may need modifications or may only apply to certain curricula. In some instances, possible modifications for individual items were documented that would allow for prompts, sample student responses, and scoring guides to reflect the different interpretations and representations of different curricula. These adaptations include small suggestions for teachers to accommodate students with different vocabulary usage (e.g. replacing the word particle with atom or vice versa) or visual representations (e.g. aligning representations of matter and energy flow by adjusting the key for a model, see Figure 3). Suggestions were also included to address differences in students' background knowledge depending on the clustering or sequencing of PEs in a given curriculum. For example, if PEs are likely to be bundled by many curricula, teachers might adapt the question or the item set such that it does or does not prompt for particular related content, or avoids redundancy with closely aligned PEs (e.g., suggestions are included for how to merge the item sets for MS-PS2-1 about collisions and MS-PS2-2 about force and mass because these PEs are highly likely to be bundled and not taught independently). Finally, a few PEs were broad enough to lead to significantly varied interpretations or emphases by different curricula, thus whole items were recommended for inclusion or exclusion depending on the curriculum of instruction. Figure 4 below shows the phenomena and content emphases for PE MS-ESS2-6 for three NGSS curricula. Each curriculum group emphasized different aspects of the PE and DCIs, and supported students to be able to explain different types of phenomena. This was also occasionally true for broader statements about the Scientific and Engineering Practices. Individual curricula sometimes emphasized different aspects of the sub-bullets of practices when several were listed (e.g., Obtaining, Evaluating, and Communicating Information highlights multiple aspects of evaluating the merit and validity of information, such as: gather, read, and synthesize information; assess credibility, accuracy, and possible bias, etc.) The incorporation of suggestions for both large and small adaptations ensures that each item set has enough individual items per PE for a fully three-dimensional assessment, even if there is an occasional item that does not work with a particular curriculum.

Figure 3: Suggestions for adapting visual representations for an item for MS-LS2-3

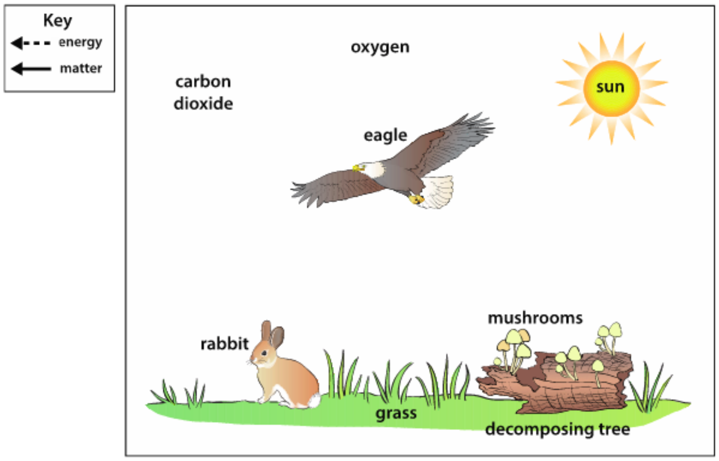
 Key energy matter oxygen carbon dioxide sun eagle mushrooms rabbit grass decomposing tree	Possible Adaptation: you may wish to adapt the key to align to representations of matter movement and energy flow used in your classroom.

Figure 4: PE and DCIs for MS-ESS2-6 and curriculum interpretations

<table>
<tr><td><b>Performance Expectation</b>

<b>MS-ESS2-6:</b> Develop and use a model to describe how unequal heating and rotation of the Earth cause patterns of atmospheric and oceanic circulation that determine regional climates. <i>[Clarification Statement: Emphasis is on how patterns vary by latitude, altitude, and geographic land distribution. Emphasis of atmospheric circulation is on the sunlight-driven latitudinal banding, the Coriolis effect, and resulting prevailing winds; emphasis of ocean circulation is on the transfer of heat by the global ocean convection cycle, which is constrained by the Coriolis effect and the outlines of continents. Examples of models can be diagrams, maps and globes, or digital representations.] [Assessment Boundary: Assessment does not include the dynamics of the Coriolis effect.]</i></td><td><b>Disciplinary Core Ideas</b>

<b>ESS2.C: The Roles of Water in Earth’s Surface Processes</b>
Variations in density due to variations in temperature and salinity drive a global pattern of interconnected ocean currents. (MS-ESS2-6)

<b>ESS2.D: Weather and Climate</b>
Weather and climate are influenced by interactions involving sunlight, the ocean, the atmosphere, ice, landforms, and living things. These interactions vary with latitude, altitude, and local and regional geography, all of which can affect oceanic and atmospheric flow patterns. (MS-ESS2-6)</td></tr>
<tr><td colspan="2"><b>Emphasis and central phenomenon for three curricula:</b>
Curriculum A
● Phenomenon: Climate change is slowing the global conveyor belt (system of ocean currents).
● Content Emphasis: the role of density in driving ocean currents
Curriculum B
● Phenomenon: The air temperature of a location has a pattern of warmer years and cooler years.
● Content Emphasis: the role of ocean currents in distributing energy and the effect on regional climates
Curriculum C
● Phenomenon: Local and global winds follow predictable patterns.
● Content Emphasis: the role of uneven heating of Earth’s surface in driving air movement</td></tr>
<tr><td colspan="2"><b>Type of adaptation:</b> The item set includes items targeted at density-driven currents, differential heating and wind, and the impact of currents on regional climate. Teachers are encouraged to only use items relevant to their curriculum.</td></tr>
</table>

In addition to the curriculum group reviews, the items were reviewed for general accessibility as a proxy for evidence of curriculum-neutrality. The expert review was implemented without consideration of any curricula, by an external team that is not associated with any curricular program. Approximately 80% of individual items rated 3 or above for the language of the prompt's phenomenon/scenario and specific question being clear and appropriate for middle school students, with a score of 3 indicating modest revisions were suggested. This aligned with our initial findings from cognitive lab interviews that students from a range of schools, including incoming 6th graders with no prior middle school curriculum experience, could engage with the items. Field test data across school sites also indicated a range of student scores for each item set. While the sample of students for a given PE was often from within the same school and therefore using the same curriculum, looking at scores across item sets and therefore across schools provides some evidence about the extent to which the design process results in items that are curriculum-neutral.

While it was not possible to complete cognitive lab interviews or field tests for all item sets due to the COVID-19 pandemic, confidence in revisions made based solely on expert review came from the strong alignment found between the expert review and the field test and cognitive lab findings from the item sets that were tested with students before the pandemic. While most items tested with students did elicit a range of responses, there were several individual items that either elicited little response, or failed to elicit evidence of understanding of one of the dimensions that was being targeted. These cases were often consistent with comments from the expert review suggesting clarification of the prompt, or more direct prompting for a particular dimension to be included with the response (as discussed above with MS-PS4-2). Figure 5 below shows researcher notes about the clarity of a prompt from a cognitive lab interview for MS-LS1-5 alongside the expert reviewer comment about that same item, and the associated revisions made to the prompt.

Figure 5: LS1-5 data about prompt clarity from cognitive lab interview and expert review

<table>
<tr><th colspan="2"><b><i>Example Item Prompt with Feedback and Revision for PE LS1-5</i></b></th></tr>
<tr><td colspan="2"><b>Initial version of prompt:</b>
Samira and her father go to the plant nursery and find 12 young snapdragon plants that are labeled “yellow flowers.” They plant them in the garden in the same kind of soil and give them the same amount of water and sunlight.
After 4 weeks, all the plants are healthy. But they are different heights and shades of yellow, as shown in the picture below. About half of them are very short and dark yellow (about 18 cm high) and the other half are tall (45 cm high) and light yellow. What might cause the differences in the plants? Explain your answer.</td></tr>
<tr><td><b>Notes from cognitive lab:</b>
<i>She [student interviewee] was confused by the prompt that stated that the “12 young snapdragon plants that are labeled ‘yellow flowers.’” She didn’t understand why these details were necessary. She thought you could either say snapdragons or yellow flowers, but saying both was confusing. She was confused that they were labeled two different ways.</i></td><td><b>Expert analysis:</b>
<i>"Consider deleting ‘snapdragon’ in the first sentence. It isn't used elsewhere, and in the rest of the description ‘yellow flowers’ is used and seems sufficient. Otherwise, it looks good."</i></td></tr>
<tr><td colspan="2"><b>Revised prompt:</b>
Samira and her father go to the plant nursery and find 12 young plants that are labeled “yellow snapdragons.” They plant them in a sunny part of the garden and give them the same amount of water.
After 4 weeks, all the plants are healthy. But they are different heights and shades of yellow, as shown in the picture below. About half of them are dark yellow and short (about 18 cm), while the other half are light yellow and tall (about 45 cm). Since all the plants grew under the same conditions, what might cause the differences in the plants? Explain your answer.</td></tr>
</table>

Patterns also emerged from our analysis of the expert review in conjunction with completed cognitive lab and field test data: if items had a multi-part prompt (e.g., complete a model and use your model to explain a phenomenon), students were likely to attend to only one part of a multi-part prompt. When not explicitly prompted, students often did not demonstrate their understanding of the crosscutting concept even if they could demonstrate this understanding in other items. These findings from student work aligned with suggestions from the expert review to break multi-part prompts, and to explicitly prompt for students to show their understanding of crosscutting concepts.

Lessons learned from item sets with complete data have allowed for improvement of the item development process and all item revision. Item development, review, and revision is ongoing.

Contribution and General Interest

In addition to the specific analysis and lessons learned described thus far, the work to develop these item sets generated important cross-curricular observations on the NGSS Middle School PEs and their assessment.

The discussions and ongoing work between different curriculum groups during the item development process has made it clear that in general it is possible to create item sets for the middle school PEs that work for teachers and students using any NGSS curriculum. However, room for interpretation of the PEs and the DCIs can sometimes lead to different treatments in curricula that might make it more difficult for students to show their understanding, and very occasionally lead to different content being taught. As discussed in the findings section, in this project these challenges were most often addressed with simple adjustments for specific vocabulary terms or visual representations in the items. Additionally, there were a small number of PEs and DCIs that were broad enough that individual curricula focused on significantly different phenomena and/or content. These differences were of a scale that in some cases the content that was focal for one curriculum may not be at all familiar to students who used a different curriculum, and unique items had to be developed to work for different learning experiences. Despite differences in curricular approaches, through the work on this project it is clear that across curriculum groups the same PEs were consistently found to be the ones where tough decisions had to be made about how to address the standard in a way that is appropriate for middle school, what aspects of the content to include, and how much of the applicable content is practical to teach within the time-constraints of a school year. These findings also aligned with discussions with the external expert reviewers who agreed with these conclusions. Across the full set of middle school PEs, in general the Physical Science PEs were most likely to have differences in vocabulary, the Life Science PEs were most likely to have differences in visual representations, and the Earth System Science PEs were most likely to have larger differences in what content was covered between curriculum groups.

In developing items for the full set of middle school PEs, it is clear that there is more content in the PEs than can be taught with integrity in the three years of middle school. Each curriculum group found this to be true independently in their curriculum development processes where decisions had to be made about where to spend instructional time. Discussion between curriculum groups also revealed that none of the curriculum programs used the evidence statements as a standard, in part because in many cases the evidence statements added content that went beyond what was already included in the PEs. Together this information raises questions about gaps that students might have in what they experience in their curriculum as compared with what they may be tested on in standardized tests. While we conclude from this work that it is possible to create three-dimensional assessments for the full set of middle school PEs, it is important to be mindful of what PEs students have been exposed to, and at what depth. Realistically, expectations may need to be tempered in terms of students' ability to demonstrate full three-dimensional understanding for all PEs.

Creating item sets for each of the PEs also highlighted some considerations related to using the PEs as written as the target for each item set. One aspect has to do with the variation in the amount or depth of content included in each PE. When compared across the full range of PEs, some PEs include DCI content that is very factual and light, and creating three-dimensional performances that kept just to that content felt forced (e.g., MS-LS1-1: Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.). Other PEs include so many complex ideas that it would take more than one class period to fully assess the PE (e.g., MS-ESS1-1: Develop and use a model of the Earth-sun-moon system to describe the cyclic patterns of lunar phases, eclipses of the sun and moon, and seasons.). Additionally, with some PEs the proscribed practice does not allow for the development of assessment items that would naturally provide evidence of student understanding of all three dimensions. For example PEs that have students obtain and evaluate information, or evaluate different designs, often need additional prompts to elicit a demonstration of understanding of the DCI or CCC content. Finally, the items sets in this project were developed to be administered on paper (or in electronic format) with no additional materials. This leads to some limitations for assessing certain practices, such as Planning and Carrying Out Investigations, and therefore the PEs that called for those practices.

The aim of this project is to design item sets to bridge the space between curriculum and curriculum-embedded assessments and standardized exams. While this goal is in sight it is critical to recognize that these are only intended to be one part of assessing a students' understanding of the PEs. The various lessons learned from this work can inform assessment development in many critical ways, and should serve as a model for researchers and practitioners developing items. In addition, as of this writing, this is the only comprehensive set of assessments for the middle school PEs, and we hope that these item sets can be used to support others both in assessing and in interpreting the PEs.

Acknowledgements

The authors would like to acknowledge the project team at the Lawrence Hall of Science for their continuing hard work, all of the teachers and students who have participated and shared their knowledge, and their colleagues at WestEd for their expert reviews and advice.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1 (5). Retrieved from http:// ejournals.bc.edu/ojs/index.php/jtla/article/view/1671

Alozie, N. M., Moje, E. B., & Krajcik, J. S. (2010). An analysis of the supports and constraints for scientific discussion in high school project‐based science. *Science Education*, *94*(3), 395–427. http://doi.org/10.1002/sce.20365

Badrinarayan, A., Wertheim, J., Penuel, B., Krajcik, J, and Smolek, T.J. (2019). Reconceptualizing Alignment for NGSS Assessment. Paper presented at NARST 2019 Annual International Conference, Baltimore, MD.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26-55.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.

Krajcik, J., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education*, *25*(2), 157–175. http://doi.org/10.1007/s10972-014-9383-2

MacPherson, A., Kastel, D., Nagle, B., Willcox, M., Jackson, W., Hariani, M., and Howarth, J. (2017). NGSS-designed assessments for a middle school ecosystems unit. Paper presented at the NARST 2017 Annual International Conference, San Antonio, TX.

McElhaney, K. W., Basu, S., Wetzel, T., and Boyce, J. (2019). Three-dimensional assessment of NGSS upper elementary engineering design performance expectations. Paper presented at NARST 2019 Annual International Conference, Baltimore, MD.

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

National Research Council (2012). *A Framework for K-12 Science Education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards, Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014). Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K–12. J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, eds. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States. 2013. *Next Generation Science Standards: For States, By State*s. Washington, DC: The National Academies Press.

Osborne, J., Pecheone, R., Quinn, H., Holthuis, N., Schultz, S., Wertheim, J., and Martin, P. (2016). A system of assessments for the next generation science standards (NGSS) in California: A discussion document. Prepared by the Stanford NGSS Assessment Project Team (SNAP). Available at https://snapgse.stanford.edu/snap-reports/snap-reports

Pellegrino, J.W., DiBello, L.V., and Goldman, S.R. (2016) A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, 51(1), 59-81.

Penuel, W.R., Van Horne, K., and Bell, P. (2016) Steps to Designing a Three-Dimensional Assessment. UW Institute for Science + Math Education. Available at stemteachingtools.org/brief/29.

Pimentel, D. S., & McNeill, K. L. (2013). Conducting talk in secondary science classrooms: Investigating instructional moves and teachers' beliefs. *Science Education*, *97*(3), 367-394.

Shepard, L. A., Penuel, W. R., & Davidson, K. L. "Design principles for new systems of assessment," *Phi Delta Kappan*, v.98, 2017.

Stanford NGSS Assessment Project (SNAP). (2018). Available at https://snapgse.stanford.edu/snap-assessments-ngss.

Zaidi, S.Z., Ko, M., Gane, B.D, Madden, K., Guar, D., and Pellegrino, J.W. (2018). Portraits of teachers using three-dimensional assessment tasks to inform instruction. Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA.