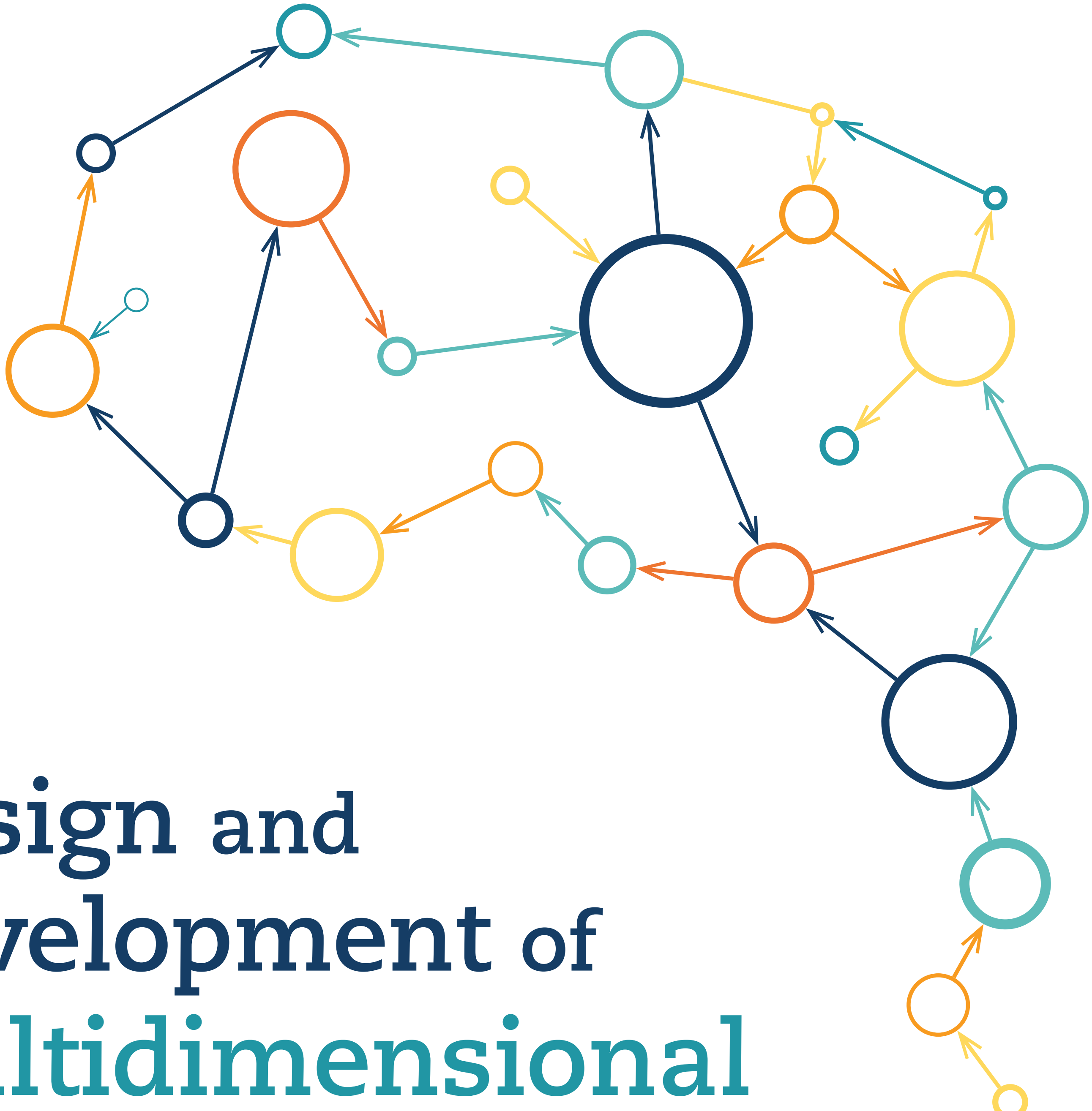


Design and Development of Multidimensional Science Item Clusters

September 2017

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

ACKNOWLEDGEMENTS

We gratefully acknowledge the Council of Chief State School Officers (CCSSO) for their role and leadership in the Science Assessment Item Collaborative (SAIC). We wish to thank the following members of the WestEd science assessment team for their expertise: Dennis Christopher, Keith Dodson, Kevin King, and Nicolle Romero.

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

WestEd
WestEd.org

The work reported herein was supported by grant number #S283B050022A between the U.S. Department of Education and WestEd with a subcontract to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The findings and opinions expressed in this publication are those of the authors and do not necessarily reflect the positions or policies of CRESST, WestEd, or the U.S. Department of Education.

This material is based upon work supported by the CCSSO SAIC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the SAIC or CCSSO.

ABSTRACT

The Next Generation Science Standards (NGSS) were designed to help students and teachers think of science not simply as content but also as how scientists think and work. In general, standards are not written as a guide to assessment but as guidance for instruction. In this paper, we examine one of the more common models that has been developed for meaningful assessment of the NGSS: a multi-item assessment known as the item cluster. We first consider each facet of an item cluster—performance expectation, phenomenon, stimulus, and item type. We then examine some of the many considerations that arise in item cluster development, such as amount of time available for assessment, depth and breadth of assessment goals, scaffolding issues for items, and scoring considerations. The paper concludes with a list of questions that should form the basis of any item cluster development process, as well as a sample development procedure for creating an item cluster.

INTRODUCTION

The Next Generation Science Standards (NGSS) are now nearly a half-decade past the initial framework document and standards development (National Research Council, 2012; NGSS Lead States, 2013), and states are now at various stages of implementing the NGSS, and other three-dimensional science standards, into classroom science curriculum and instructional practices. The primary goals of the NGSS are to develop a conceptual understanding of scientific content by means of using scientific and engineering practices (SEPs) to elicit this understanding, and to emphasize how underlying scientific crosscutting concepts (CCCs) help tie the disciplinary core ideas (DCIs) together.

As numerous states have adopted, or adapted, the NGSS, a great deal of interest and discussion has focused on the question of how to most effectively assess student proficiency, not only on the NGSS content (DCIs), but also on the related SEPs and CCCs that together comprise performance expectations (PEs) that form the basis of the NGSS (National Research Council, 2012; Harris, Krajcik, Pellegrino, & McElhaney, 2016). The three dimensions of the NGSS (SEPs, DCIs, CCCs) provide great opportunities to effectively change how students learn science. The goal of these new standards is to set a new model in which science instruction moved away from memorization of theories and facts toward a deeper engagement with scientific thinking and practices. The initial documents did not emphasize how student proficiency with these new standards would be assessed, and only later did papers emerge outlining the challenges of NGSS assessment (National Research Council, 2014).

In 2015, WestEd, in conjunction with CCSSO and the Science Assessment Item Collaborative (SAIC), developed an Assessment Framework for NGSS-based testing (CCSSO, 2015a). A major conclusion of that document was that full assessment of the three dimensions of each PE requires an approach more robust than traditional single-item testing. To fully assess each PE, multiple test items, each tied to a unifying context, were necessary in order to build a testing framework that lent support to assessment of the SEPs and CCCs. A companion document (CCSSO, 2015b) to the NGSS Assessment Framework provided Item Specifications Guidelines that discussed basic considerations that inform NGSS item development, and also provided a sample set of item specifications for an item cluster (IC). This sample set included mappings showing connections between the newly released NGSS Evidence Statements (Next Generation Science Standards Lead States, 2015) and dimensions specific to the PEs. Item cluster prototypes for grade 5 and high school were subsequently developed by WestEd in collaboration with SAIC (CCSSO, 2015a).

The item cluster does not represent a fixed template. Variations of the basic idea of an IC have been introduced and developed by several state educational agencies and, at the local level, by instructors developing formative assessments for their classroom. The idea that the PEs contained in the NGSS can most fully be assessed by utilizing several items that align to different combinations of dimensions specific to PEs is the underlying tenet of all item cluster variations. Another idea that most ICs try to incorporate is basing the IC around a stimulus (based on an identified phenomenon) that is amenable to scientific investigation, thus tying into the NGSS philosophy of learning science by doing science and thinking scientifically.

As is generally the case when any new process is developed, the development of the item cluster prototypes reinforced some early guiding ideas but also provided many unanticipated lessons learned. In this document, we discuss the item cluster development process, focusing on key conclusions and guiding principles that have been laid out and subsequently refined through ongoing IC development efforts.

This paper is one of three companion documents discussing issues related to the development of assessment materials aligned to the NGSS. For information on issues around proper alignment to the NGSS, please see the companion paper "Alignment Considerations for Next Generation Science Standard Assessments," (CSAI, 2017a). For information on processes for maintaining quality while scaling up the development of item clusters, please see the companion paper "Quality Expectations and Development Considerations of Item Clusters Assessing Multidimensional Science Standards," (CSAI, 2017b).

THE ITEM CLUSTER DEVELOPMENT PROCESS

Key Considerations

Numerous planning and design decisions must be made when developing an item cluster, which is a key reason why much front-end planning is essential in IC development. Some of the fundamental elements of an item cluster are listed in Table 1. For each of these elements, there are several questions that must be considered prior to beginning development of an item cluster.

Table 1: **Item Cluster Development**

Item Cluster Element	Driving Questions	Options
Performance Expectations	How many PEs should be assessed in a single IC? Is assessment across multiple DCI domains in a single IC allowable and/or desirable? What role should the Evidence Statements take?	1. Single-domain, single-PE IC 2. Single-domain, multi-PE (see **Figure 1**) 3. Multi-domain, multi-PE
Phenomenon	What should be the observable, natural process or occurrence that forms the foundation for the ensuing IC items?	Numerous, but all phenomena should be robust enough to support multiple items, be grade-level appropriate, and be stated as an observable event that science knowledge and investigation can be used to explain or predict.
Stimulus	How should the supporting text for the IC be distributed among the items? What information or data should students be given to support assessment of the DCI, SEP, and CCC dimensions?	1. Multiple stimuli of roughly comparable length preceding most items within the cluster. 2. An initial, longer stimulus providing enough context to support several items, followed by a few shorter stimuli to support later items. 3. A single lengthy stimulus with supporting stimuli within the items if needed. All of the items within an IC should relate to the overarching stimuli.
Item Types	What mix of item types will be available within the IC? Which item types are most appropriate for assessing specific SEPs and CCCs?	Numerous types to choose from, with most common being multiple choice, multiple select, evidence-based selected response, constructed response, and technology-enhanced items (TEIs).

In addition, the choices made about the IC elements listed in Table 1 must be balanced in concert with specific assessment criteria and limitations. Among these are time (how long should it take an average student to complete an IC?), breadth of coverage (what subset of PEs are to be covered in an assessment?), testing capabilities (will computers be available to all students?), and assessment goal (formative or summative?).

In the sections that follow, each of the elements listed in Table 1 are addressed in more detail, with examples provided to illustrate how they might be addressed in the development of an actual IC.

Performance Expectations (PEs)

One of the first decisions that needs to be made in developing ICs is determining the appropriate breadth and depth of the material to be assessed by each IC. Among the numerous PEs in each grade band, some dovetail very naturally with other PEs and may share one or two of the same dimensions. The DCIs may also vary significantly in their scope of content. Some PEs have a narrowly focused DCI, centered on one main idea, while other PEs are broad in scope and encompass several related concepts within their DCI. The latter type of PEs are often robust enough to support an IC by themselves. Recognition of the similarities and variations between the PEs in terms of content, scope, and dimensional alignment must be brought to bear in determining whether a PE should be assessed independently in an IC or as a bundle consisting of one or more PEs.

In practice, experience has shown that three is the maximum number of PEs that can be effectively assessed in a single IC. Having more than three PEs requires that some PEs only be assessed in an incomplete or superficial manner for a typical IC. Most of the ICs developed by WestEd have included a maximum of two PEs, although there are some cases in which three PEs work well together and invite the development of a multi-PE IC. Most of these ICs bundle PEs from the same DCI domain (i.e., life science, physical science, earth and space science), although there are some cases in which PEs from different domains share conceptual linkages that allow them to be joined together within a single IC. While multi-PE ICs generally overlap in at least one dimension (often sharing the same CCC), there are certainly situations in which there is no dimensional overlap between the component PEs. An example in the grades 3–5 span are PEs 3-LS4-1 *(Analyze and interpret data from fossils to provide evidence of the organisms and the environments in which they lived long ago.)* and 4-ESS1-1 *(Identify evidence from patterns in rock formations and fossils in rock layers to support an explanation for changes in a landscape over time.)*.

Figure 1 illustrates the dimensional alignment of a single-domain multi-PE IC in which both the DCI (Scale and Properties of Matter) and the CCC are common among the two PEs in the bundle. Figure 2, in turn, shows the dimensional alignment of the multi-domain IC consisting of PEs 3-LS4-1 and 4-ESS1-1.

Figure 1: Item Cluster with Two PEs from the Same Domain

Level:	Grade 5
Primary Target Domain:	Physical Science
Target PEs:	5-PS1-1, 5-PS1-2
Crosscutting Concept(s) Focus:	Scale, Proportion, and Quantity
Science and Engineering Practice(s) Focus:	Developing and Using Models, Using Mathematics and Computational Thinking
Reasoning for PE Groupings:	Mass (size micro to macro), and conservation of mass
Phenomenon:	Sugar is no longer visible when it dissolves in water, but the mass of the mixture stays the same
Allowable Item Types:	SR, TE, CR

	5-PS1-1	**5-PS1-2**
Performance Expectations:	Develop a model to describe that matter is made of particles too small to be seen.	Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling, or mixing substances, the total weight of matter is conserved.
Target Clarifications:	Examples of evidence could include adding air to expand a basketball, compressing air in a syringe, dissolving sugar in water, and evaporating salt water.	Examples of reactions or changes could include phase changes, dissolving, and mixing that form new substances.
Assessment Boundary:	Assessment does not include the atomic-scale mechanism of evaporation and condensation or defining the unseen particles.	Assessment does not include distinguishing mass and weight.
Disciplinary Core Idea(s):	**PS1.A: Structure and Properties of Matter** • Matter of any type can be subdivided into particles that are too small to see, but even then, the matter still exists and can be detected by other means. A model showing that gases are made from matter particles that are too small to see and are moving freely around in space can explain many observations, including the inflation and shape of a balloon and the effects of air on larger particles or objects.	**PS1.A: Structure and Properties of Matter** • The amount (weight) of matter is conserved when it changes form, even in transitions in which it seems to vanish. **PS1.B: Chemical Reactions** • No matter what reaction or change in properties occurs, the total weight of the substances does not change. (Boundary: Mass and weight are not distinguished at this grade level.)
Science and Engineering Practice(s):	**Developing and Using Models** Modeling in 3–5 builds on K–2 experiences and progresses to building and revising simple models and using models to represent events and design solutions. • Use models to describe phenomena.	**Using Mathematics and Computational Thinking** Mathematical and computational thinking in 3–5 builds on K–2 experiences and progresses to extending quantitative measurements to a variety of physical properties and using computation and mathematics to analyze data and compare alternative design solutions. • Measure and graph quantities such as weight to address scientific and engineering questions and problems.
Crosscutting Concept(s):	**Scale, Proportion, and Quantity** • Natural objects exist from the very small to the immensely large.	**Scale, Proportion, and Quantity** • Standard units are used to measure and describe physical quantities such as weight, time, temperature, and volume. ***Connections to Nature of Science*** **Scientific Knowledge Assumes an Order and Consistency in Natural Systems** • Science assumes consistent patterns in natural systems.

Figure 2: Item Cluster with Two PEs from Different Domains

Level:	Grade 4
Primary Target Domain:	Connection between Life Science and Earth and Space Science
Target PEs:	3-LS4-1 and 4-ESS1-1
Crosscutting Concept(s) Focus:	Scale, Proportion, and Quantity and Patterns
Science and Engineering Practice(s) Focus:	Analyzing and Interpreting Data and Constructing Explanations and Designing Solutions
Reasoning for PE Groupings:	Patterns in rock layers and fossils
Phenomena:	Fossils on one side of the Earth are in the same rock structure on the other side of the Earth
Allowable Item Types:	SR, TE, CR

	3-LS4-1	**4-ESS1-1**
Performance Expectations:	Analyze and interpret data from fossils to provide evidence of the organisms and the environments in which they lived long ago.	Identify evidence from patterns in rock formations and fossils in rock layers to support an explanation for changes in a landscape over time.
Target Clarifications:	Clarification Statement: Examples of data could include type, size, and distributions of fossil organisms. Examples of fossils and environments could include marine fossils found on dry land, tropical plant fossils found in Arctic areas, and fossils of extinct organisms.	Examples of evidence from patterns could include rock layers with marine shell fossils above rock layers with plant fossils and no shells, indicating a change from land to water over time; and, a canyon with different rock layers in the walls and a river in the bottom, indicating that over time a river cut through the rock.
Assessment Boundary:	Assessment does not include identification of specific fossils or present plants and animals. Assessment is limited to major fossil types and relative ages.	Assessment does not include specific knowledge of the mechanism of rock formation or memorization of specific rock formations and layers. Assessment is limited to relative time.

Disciplinary Core Idea(s):	**LS4.A: Evidence of Common Ancestry and Diversity**	**ESS1.C: The History of Planet Earth**
	• Some kinds of plants and animals that once lived on Earth are no longer found anywhere. (*Note: moved from K-2*) • Fossils provide evidence about the types of organisms that lived long ago and also about the nature of their environments.	• Local, regional, and global patterns of rock formations reveal changes over time due to earth forces, such as earthquakes. The presence and location of certain fossil types indicate the order in which rock layers were formed. (4-ESS1-1)
Science and Engineering Practice(s):	**Analyzing and Interpreting Data**	**Constructing Explanations and Designing Solutions**
	Analyzing data in 3–5 builds on K–2 experiences and progresses to introducing quantitative approaches to collecting data and conducting multiple trials of qualitative observations. When possible and feasible, digital tools should be used. • Analyze and interpret data to make sense of phenomena using logical reasoning.	Constructing explanations and designing solutions in 3–5 builds on K–2 experiences and progresses to the use of evidence in constructing explanations that specify variables that describe and predict phenomena and in designing multiple solutions to design problems. • Identify the evidence that supports particular points in an explanation. (4-ESS1-1)
Crosscutting Concept(s):	**Scale, Proportion, and Quantity**	**Patterns**
	• Observable phenomena exist from very short to very long-time periods. *Connections to Nature of Science* **Scientific Knowledge Assumes an Order and Consistency in Natural Systems**	• Patterns can be used as evidence to support an explanation. (4-ESS1-1) *Connections to Nature of Science* **Scientific Knowledge Assumes an Order and Consistency in Natural Systems**
	• Science assumes consistent patterns in natural systems.	• Science assumes consistent patterns in natural systems. (4-ESS1-1)

While each PE in the NGSS is a composite of three dimensions, developing a single item that fully encapsulates student understanding of the CCC, SEP, and DCI of a PE is very challenging. For this reason, typically ICs have been developed around the philosophy that each item in an IC should align to at least two dimensions, and that the IC as a whole should align to all three dimensions (see CSAI (2017a) for a deeper discussion of alignment issues between assessment items and the dimensions of the NGSS PEs).

Items may also consist of multiple parts. A typical example would be a two-part item in which the first part asks the student to make an identification or select an option, and the second part requires the student to supply the reasoning for their choice in the first part. Such two-part items are becoming increasingly common in NGSS assessment and are frequently collectively referred to as "evidence-based selected response" (EBSR) item types.

Choosing PEs for bundles also sometimes involves scaffolding considerations, especially in grades K–5. For example, PEs in grade 5 frequently build off DCIs covered in PEs from earlier grades. Thus, when choosing to bundle two related PEs from different grades, the IC for the bundle will generally need to follow the same progression of ideas as the PEs themselves.

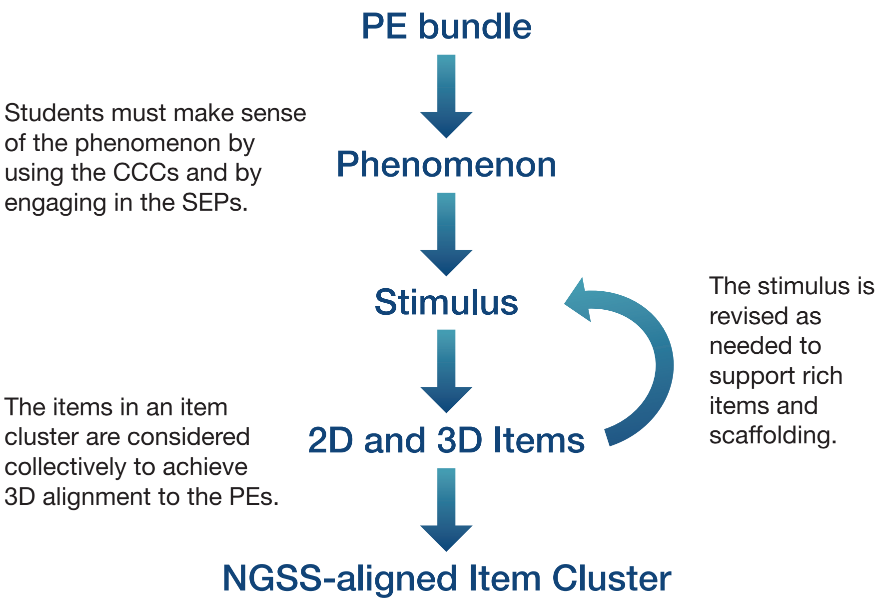
Role of Evidence Statements: As noted in the Introduction, each PE is supported by an evidence statement (Next Generation Science Standards Lead States, 2015) that provides acceptable evidence of understanding and application of the PE. As a tool for IC development, the evidence statements can aid in identifying the pieces of observable evidence that demonstrate student mastery of the PE. The organization of an evidence statement for a PE is driven by the PE's SEP. Thus, for example, evidence statements for PEs having the modeling SEP will be focused on the component pieces of a model to be created by the student and how the model can subsequently be used to flesh out deeper ideas and connections (i.e., the DCI and CCC aspects of the PE).

While the evidence statements are certainly useful in developing ICs, they do not necessarily provide a road map for item cluster development. Many of the evidence statements are quite long, covering a wide swath of scientific ideas that would prove unwieldy to fully assess within a single IC. Given this situation, it is wise to recall that assessment alignment to a PE is tied to the three dimensions of the PE, and not to the enumerations listed in the PE's evidence statement. Most ICs will certainly align to some, or even most, of the evidence statements for a PE. Ultimately, however, it is often not practical to rigidly adhere to the PE's evidence statement because in order to encompass the entire evidence statement, one runs the risk of having to anchor the IC to a phenomenon that is much too broad in scope. The resulting IC would therefore be diffuse, rather than focused, and longer than is desirable.

Phenomenon

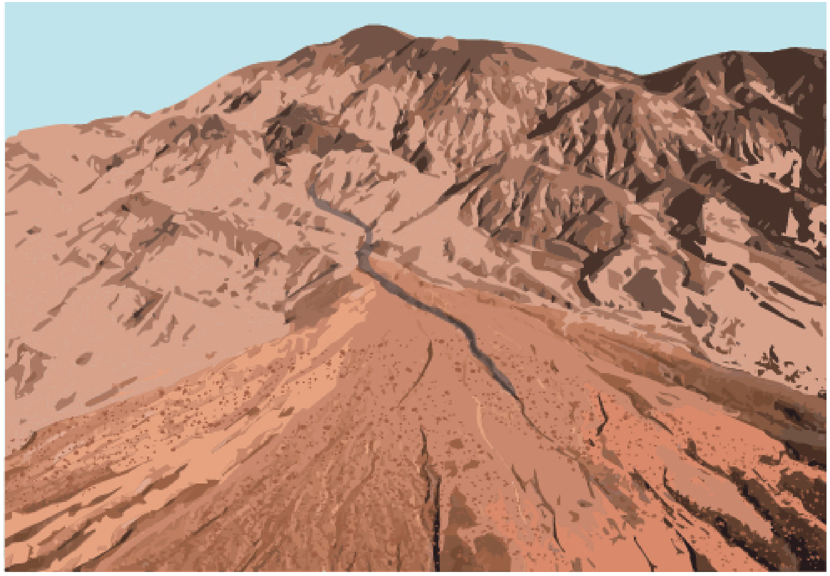
Frequently, the success of an IC in demonstrating student proficiency of a PE or bundle of PEs may depend most strongly on the richness of the underlying phenomenon that serves as the inquiry foundation of the IC (see Figure 3). This is not to say that the phenomena used for ICs must be based on complex observations that immediately require the student to perform multiple levels of analysis in order to engage with the underlying scientific ideas. Indeed, such phenomena are frequently prone to going beyond the boundaries of the DCI dimensions of PEs.

Figure 3: **Role of Phenomena in Item Clusters**



Observable phenomena do not need to be complex, but should provide windows of opportunity for various levels of scientific explanation and exploration. A relatively simple observation—for example, the characteristic triangle shape of an alluvial fan emerging from the foothills of a steep mountain slope—can serve as the basis of a rich IC on erosion and deposition processes (see Figure 4).

Figure 4: **Graphic of an Alluvial Fan**



Ideally, a phenomenon should provide opportunities to assess student understanding at increasing levels of complexity. Content scaffolding is necessary in many ICs in order to fully assess all three dimensions, making a phenomenon that lends itself to progressively more complex explanations based on natural extensions of the original phenomenon highly desirable. Returning to the alluvial fan example, such an extension might be asking the student to explain how (perhaps by reference to a previously developed model) the pattern of stream channels on the alluvial fan provides information about how the alluvial fan formed.

There is a feedback relationship between PEs and phenomena that is often observed early on in item cluster planning. In particular, if a PE suggests only one or two natural phenomena, that is an indication that the IC is too narrowly defined by the single PE. Consequently, in such an event, adding a PE to create a PE bundle is often more appropriate in order to broaden the types of phenomena that could be used for the IC.

Stimuli and Scaffolding Issues

The stimuli for an IC consist of all the textual material and information not included in the item prompts or the items themselves. The initial stimulus identifies the phenomenon for the IC, any necessary background information or data needed for at least the initial item(s), and the associated context of the IC, if any.

One occasional point of confusion is the difference between a context and a phenomenon. At the most basic level, a context is the background setting for the study of a phenomenon. Returning once more to the alluvial fan example, a context for an IC on this phenomenon might be a group of students on a field trip to the desert observing natural features of the desert terrain. The context (storyline) can be woven throughout the IC, but the assessment components of the IC are always based on the phenomenon as it connects to the PE and associated dimensions. Generally, IC contexts are relatively thin, as they are not the basis of assessment and can add to the student's reading load if too detailed. In fact, ICs do not absolutely require a context, but they are generally included to better connect the items within the IC.

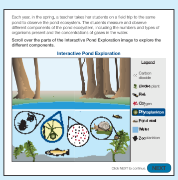
The stimuli provide the framework for the IC items and must be carefully planned when developing the structure of the IC. Unlike standalone test items, the entire stimulus need not precede the first item in an IC. In fact, because of scaffolding needs, it is generally best not to present too much information initially as it can cue the answers to early items in the IC. Assuming that the assessment is given online, the IC can be designed in such a way that students cannot go back to previously answered items when they are given information in later IC stimuli that might clue the answer to earlier items (e.g., item blocking). In addition, because incorrect responses on initial IC items can affect a student's ability to correctly answer subsequent items due to the scaffolded structure of an IC, it is sometimes desirable to provide the correct responses to earlier items in later stimuli so that a student is not penalized for his or her initial incorrect responses. An alternative approach to addressing the scaffolding issue is to supply new sample data in subsequent items so that any erroneous data from a student's earlier responses are not carried over.

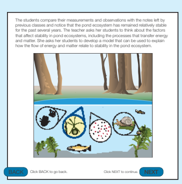
WestEd's experience has been that a typical IC with four or five items will generally work best with at least two stimuli sections and frequently three. The initial stimulus is usually the longest, in part because it introduces both the phenomenon and the context. As shown in Figure 3, the stimuli and their associated items are developed together and go through numerous iterations in the development process to ensure that they are tightly woven together and supply the scaffolding framework necessary to support subsequent stimuli and items without causing cognitive overload.

Table 2 shows how stimuli and items relate to each other in the high school SAIC IC prototype (CCSSO, 2015). Although not shown in the table, each of the three stimuli also include interactive graphics. Stimulus 3 includes the same graph as the one in stimulus 2, but with the addition of a dotted line to the graph to help support the scaffolding of the ideas assessed by item 5. Most stimuli for any IC will include at least one graphic or table, whether interactive or static. Note that stimulus 1 is significantly longer than the other two stimuli and supports two items having a combined four parts. The PE bundle for this sample IC consisted of two PEs that were tied to the Develop and Use Models SEP. Thus, all the stimuli refer to models, which are either created by the student through the use of labeling (stimulus 1) or are given to the student (stimuli 2 and 3) and then used for reasoning-based items.

Table 2: **Relationship Between Stimuli and Items in a Typical IC** (Continued on the next page...)

Stimulus 1



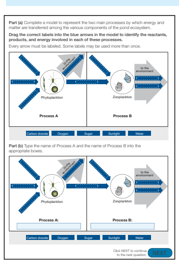


Each year, in the spring, a teacher takes her students on a field trip to the same pond to observe the pond ecosystem. The students measure and observe different components of the pond ecosystem, including the numbers and types of organisms present, the temperature of the water and concentrations of gases in the water.

Scroll over the parts of the Interactive Pond Exploration image to explore the different components.

The students compare their measurements and observations with the notes left by previous classes and notice that the pond ecosystem has remained relatively stable for the past several years. The teacher asks her students to think about the factors that affect stability in pond ecosystems, including the processes that transfer energy and matter. She asks her students to develop a model that can be used to explain how the flow of energy and matter relate to stability in the pond ecosystem.

Item 1

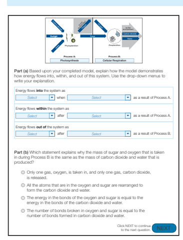


Part (a) Complete a model to represent the two main processes by which energy and matter are transferred among the various components of the pond ecosystem.

Drag the correct labels into the blue arrows in the model to identify the reactants, products, and energy involved in each of these processes.

Part (b) Type the name of Process A and the name of Process B into the appropriate boxes.

Item 2

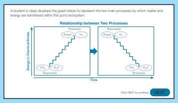


Part (a) Based upon your completed model, explain how the model demonstrates how energy flows into, within, and out of this system. Use the drop-down menus to write your explanation.

Part (b) Which statement explains why the mass of sugar and oxygen that is taken in during Process B is the same as the mass of carbon dioxide and water that is produced?

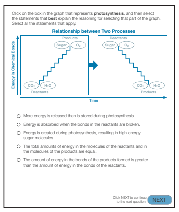
Table 2: **Relationship Between Stimuli and Items in a Typical IC** (...continued from previous page)

Stimulus 2



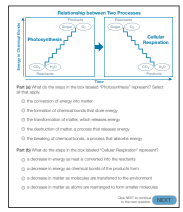
A student in class develops the graph below to represent the two main processes by which matter and energy are transferred within this pond ecosystem.

Item 3



Click on the box in the graph that represents photosynthesis, and then select the statements that best explain the reasoning for selecting that part of the graph. Select all the statements that apply.

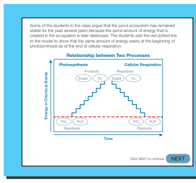
Item 4



Part (a) What do the steps in the box labeled Photosynthesis” represent? Select all that apply.

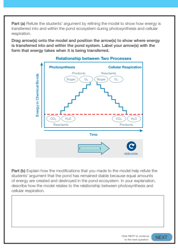
Part (b) What do the steps in the box labeled Cellular Respiration” represent? Select all that apply.

Stimulus 3



Some of the students in the class argue that the pond ecosystem has remained stable for the past several years because the same amount of energy that is created in the ecosystem is later destroyed. The students add the red dotted line to the model to show that the same amount of energy exists at the beginning of photosynthesis as at the end of cellular respiration.

Item 5



Part (a) Refute the students' argument by refining the model to show how energy is transferred into and within the pond ecosystem during photosynthesis and cellular respiration.

Drag arrow(s) onto the model and position the arrow(s) to show where energy is transferred into and within the pond system.

Part (b) Explain how the modifications that you made to the model help refute the students' argument that the pond has remained stable because equal amounts of energy are created and destroyed in the pond ecosystem. In your explanation, describe how the model relates to the relationship between photosynthesis and cellular respiration.

Item Types

Several questions should be addressed before selecting the item types for use in ICs. Among these are:

- Is the assessment for summative or formative purposes?
- Will the assessment be machine scored, hand scored, or a mix of both?
- If the assessment is delivered online, what are the available types of technology-enhanced item templates?
- What are the time constraints for the assessment?

Ideally, an item cluster should involve a number of different item types, as some item types work better for certain types of questions. For example, PEs that ask students to develop a model are generally more difficult to assess, especially for online, machine-scored tests. It is not reasonable for students to develop a unique model on an assessment, so most test questions tied to a modeling SEP have involved students producing all or parts of a pre-existing model. Technology-enhanced (TE) items that involve students moving labels or arrows to their appropriate spots within a model framework have been very useful in aligning items to the modeling SEP. For more information on available item types, see the companion paper "Quality Expectations and Development Considerations of Item Clusters Assessing Multidimensional Science Standards" (CSAI, 2017b).

Time Considerations: When developing ICs, there is usually a targeted amount of time that each IC should take to complete. Each item type requires a different amount of time to answer. Multiple choice (MC) items will usually require the least amount of time, (about 1 minute); TE items vary based on the specific type but on average require more time than MC items; and constructed response (CR) items require the most amount of time of any item type because of the writing expectation. For this reason, our development experience has been to avoid including more than one CR in a given IC unless a significant amount of time has been allotted for the student to complete the IC.

Scoring Considerations: Multiple-choice tests are the simplest to administer and the easiest to score, but a single multiple-choice question is limited in its ability to fully assess a PE. For this reason, ICs have generally included a mix of the three item types (MC, TE, and CR). However, CR items and TE items often involve either more complex reasoning or more individual decisions in determining the correct answer(s). For example, random guessing on a multiple-choice question gives a student a 25% chance of choosing the correct answer, but such a strategy on a Multiple Choice Dynamic question with three dropdown menus having four choices each drops the guesser's likelihood of a correct answer to 1.6%. Even a multiple select question with five options and two correct answers reduces the guesser's odds of a correct answer to 10%. Similarly, a CR question in which a student has to explain a process and justify his or her reasoning involves multiple steps and may necessitate two or more points with partial credit rubrics.

Putting It All Together

As the previous sections illustrated, there are many pieces to an item cluster and numerous practical development decisions that must be tracked. To summarize, keep the following questions in mind:

- Are all the dimensions specific to the PEs in a PE bundle aligned to at least one item?
- Does every item in the IC align to at least two dimensions of a PE?
- Do the stimuli tie the IC together and stay focused on the central phenomenon?
- How many points should each item receive?
- What is the estimated time it will take a student to complete the IC?
- What is the distribution of item types in the IC?

Given the many moving parts and basic requirements of an IC, WestEd developers have found it very useful to track the various IC elements and their structure using a metadata grid in a spreadsheet. Table 3 shows such a grid corresponding to the two-PE IC introduced in Figure 1. Metadata for individual items within the IC can also be tracked using item templates corresponding to specific item types (Figure 5).

Table 3: **Metadata Tracking Table Used in Item Cluster Development**

Item	Item Part	Brief Description	Item Type	PE	DCI	SEP	CCC	EV Level	EVs	Points	Estimated Time (min)	Hand or Automated Scoring
Stimulus		Preparing lemonade	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3	N/A
1	1	Designing and populating a data table	Text Entry/ Table Fill-In	5-PS1-2	N/A	5	3	1	1.a.i 1.a.ii	2	2	A
2	2a	Calculate mass of ingredient	Computation	5-PS1-2	PS1.A PS1.B	5	3	1	1.a.i 1.a.ii	1	1	A
	2b	Graphing masses of ingredients	Graphing					2	2.a	2	2	A
	2c	Describe properties of individual ingredients	Short Answer					2	2.c	1	2	H
3	3a	Claim for conservation of mass	Multiple Choice	5-PS1-2	PS1.A PS1.B	5	3	2	2.d	1	1	A
	3b	Identify evidence of conservation of mass	Multiple Select					2	2.d	1	1	A
Stimulus		Investigating ingredients	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1	N/A
4	4a	Describe that both sugar and water are made up of particles	Short Answer	5-PS1-1	PS1.A	2	3	1	1.a.ii	1	2	H
	4b	Building a model to show particles of matter	Building a Model (Drag-and-Drop)					1	1.a.i 1.a.ii	1	3	A or H
5	5a–b	Describing the model and use of model in explaining science phenomenon	Constructed Response	5-PS1-1	PS1.A	2	3	2, 3	2.a.i 3.a	2	6	H
								Total:	**9 of 11**	**12**	**24**	

Figure 5: Item Template for a Multiple Choice or Multiple Select Item

Item 1

Alignment: ***[****PE]* (SEP / DCI / CCC) *[add colon + additional alignment(s) if applicable]*
(Evidence Statements: *[EV statement code(s)]*)
Item Type: *[type]* (Part A), *[type]* (Part B)
Points:
NOTE: Item 1 is blocked after it is answered.

Part A

[Stem.]

[graphic(s), if applicable]

KEY *[if applicable]*
[Text or graphic rubric.]

Part B

[Stem.]

- ○ A.
- ○ B.
- ○ C.
- ○ D. *
- ○ E. *[MS]*
- ○ F. *[MS]*
- ○ G. *[MS]**
- ○ H. *[MS]*

[=correct response(s)]*

Item Cluster Development Approach

The individual steps in IC development can vary. The list below indicates how WestEd, through IC development experience, has laid out a step-by-step development procedure.

1. Identify the PEs that are to be assessed, the duration of testing, and the natural PE bundles within the group of PEs to be assessed.
2. If the duration is too short to develop ICs assessing the entire initial list of PEs to be assessed, consider potential alternatives (e.g., define a subset of the initial PEs to be assessed, eliminate item types such as CR items that require more time to answer).

3. Define the phenomenon that will be the subject of the IC. Note that this step sometimes requires a rethinking and possible reorganizing of the underlying PE bundles.
4. Lay out the plan for the items to be assessed and the necessary stimulus material to serve as the basis for these items. As items are brainstormed, thought should also be given to what item types will work best to elicit evidence of student understanding and to align to two or more dimensions specific to a PE.
5. Review the plan with all stakeholders and continue to revise as needed.
6. Write the IC after the plan has been agreed upon and begin initial art development.
7. Edit the IC, focusing on clear alignment, clarity, completeness, strong connections between items without cueing, and conformity to time considerations.

REFERENCES

Council of Chief State School Officers (CCSSO), Science Assessment Item Collaborative. (2015a). Assessment Framework for the Next Generation Science Standards. Washington, DC: Council of Chief State School Officers.

Council of Chief State School Officers (CCSSO), Science Assessment Item Collaborative. (2015b). Assessment Framework for the Next Generation Science Standards. Washington, DC: Council of Chief State School Officers.

The Center on Standards and Assessment Implementation (CSAI). (2017a). *Alignment Considerations for NGSS Assessments*. San Francisco, CA: The Center on Standards and Assessment Implementation.

The Center on Standards and Assessment Implementation (CSAI). (2017b). *Quality Expectations and Development Considerations of Item Clusters Assessing Multidimensional Science Standards*. San Francisco, CA: The Center on Standards and Assessment Implementation.

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.

National Research Council. (2012). A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: The National Academies.

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

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

Next Generation Science Standards Lead States. (2015). Evidence Statements.
Available at https://www.nextgenscience.org/evidence-statements