Quality Expectations and Development Considerations of Item Clusters Assessing Multidimensional Science Standards

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: Meghan Bell, Raymond de Hont, Kevin King, and Nicolle Romero.

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

WestEd supports multiple states in the development of their next generation science assessments. Through this work, a number of effective approaches and strategies related to the measurement of multidimensional science standards have emerged, and several key considerations can be highlighted to support similar efforts in designing and developing effective next generation science assessments. One innovative approach to assessing three-dimensional standards that has emerged is the use of item clusters, such as those presented in the SAIC Assessment Framework. This paper explores quality expectations and development considerations of effective item cluster design and authoring.

INTRODUCTION

The Next Generation Science Standards (NGSS), which integrate three dimensions that are reflective of how science and engineering are practiced in the real world, were released to states in 2013. The integration of the three dimensions—Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs)—into three-dimensional Performance Expectations (PEs) requires innovative approaches to building a multidimensional science assessment.

One innovative approach to assessing three-dimensional standards that has emerged is the use of item clusters. NGSS-aligned item cluster prototypes, as well as the Assessment Framework and Item Specifications Guidelines documents, were developed by WestEd in collaboration with the Council of Chief State School Officers (CCSSO) and participating states as part of the Science Assessment Item Collaborative (SAIC) in 2016. The item cluster design was based on the recommendations that NGSS assessment tasks should contain multiple components, such as a set of interrelated questions, in order to appropriately measure the three intertwined dimensions of the NGSS (NRC, 2014).

An item cluster consists of at least one stimulus and a set of four to six interrelated questions that follow a storyline, that provide students the opportunity to use their knowledge of the three dimensions to make sense of a phenomenon in a logical and scaffolded manner. Item clusters allow students to engage with a phenomenon, as well as provide the much-needed depth of the standards that is often missing from more traditional science assessments. For information on issues around proper alignment to the NGSS, please see the companion paper "Alignment Considerations for NGSS Assessments," (CSAI, 2017a). For information on the process of developing an NGSS-aligned item cluster, please see the companion paper "Design and Development of Multidimensional Science Item Clusters," (CSAI, 2017b).

Each item cluster is developed to align to 1–3 PEs that are bundled together to support the full exploration of a phenomenon using all three dimensions. Each item within the item cluster must assess at least two of the three dimensions, and is structured in a way to assimilate information provided in the stimulus. When the item cluster is considered holistically, the entire item cluster assesses and aligns to, at minimum, one SEP, one DCI, and one CCC. A detailed description of NGSS assessment alignment expectations is provided in a companion paper titled "Alignment Considerations for NGSS Assessments."

To support the deep exploration of a phenomenon, each item cluster requires a significant investment of testing time and a high degree of cognitive load on the part of the student. Consequently, the total number of item clusters that can be included in an assessment is limited by caps on total testing time, and may be influenced by other factors, such as student fatigue, development costs, scoring considerations, and platform constraints. The design of the assessment must also be deliberate, and consideration must be given to the number and length of each item cluster, the overall desired breadth of PEs on the assessment, the available item types, methods of scoring, delivery platform, grade level, and development timelines and costs.

ITEM CLUSTERS AND STANDALONES AS MEASUREMENT TOOLS OF THE NGSS

Most large-scale summative assessments are constrained by several factors, including, but not limited to, the total amount of testing time for a given administration. Due to the inherent demands of the item cluster (testing time and student cognitive load), it may not be feasible to measure the desired breadth of the standards utilizing item clusters alone on a summative assessment. One option to consider is the use of standalone items in combination with item clusters. Following the guidelines set forth by the Board on Testing and Assessment that no dimension should be assessed in isolation (NRC, 2014), every item, including standalones, should align to at least two of the three dimensions. While standalone items do not allow for the deep exploration of phenomena as do item clusters, they do provide additional data and opportunities to gather evidence of student proficiency relative to all three dimensions and a broader range of NGSS PEs than just those in the item clusters on the assessment.

KEY CONSIDERATIONS WHEN DEVELOPING ITEM CLUSTERS FOR SUMMATIVE ASSESSMENTS

Item clusters provide an opportunity for students to delve deeply into a specific phenomenon through the use of targeted practices, content knowledge, and crosscutting concepts associated with a selected

PE or bundles of PEs. The following considerations that should be addressed before or during the development of item clusters will be explored:

- Item Type Selection and Platform Functionality
- Establishing Clear Alignment Expectations
- Training: Moving Beyond Traditional Science Assessment
- Developing Guiding Documentation
- Developing Item Clusters at Scale

Item Type Selection and Platform Functionality – The availability of effective item types used to measure proficiency relative to the NGSS PEs should be determined when evaluating the standards targeted for an assessment. The range and robustness of the suite of item types differ somewhat from platform to platform and will influence how a subset of available item types can be leveraged to best assess all three dimensions of the NGSS.

Specific item types, such as technology-enhanced items (TEI), can be used to better assess certain dimensions. For example, TEIs that allow students to manipulate data can be leveraged to more effectively gather evidence of student proficiency with respect to certain SEPs than other item types. Constructed response (CR) items, both single-prompt and multi-part, are especially useful for developing items that align to all three dimensions (DCI, SEP, and CCC). Table 1 below provides information on how select item types lend themselves to targeting specific dimensions more effectively than others.

Table 1: Effective Item Types for Assessing Different Combinations of Dimensions

Item Type	Targeted Dimensions	Description
Drag-and-drop	SEP and DCI; DCI and CCC	Draggers can be text or images used to complete a model or sort statements, classify objects, etc. Good option for SEP 2 Developing and Using Models.
Graphing (Bar, Line)	SEP (4 and 5) and DCI	Students drag bars/lines to indicate values or trends of real or predicted data.
Selected Response (MC, MS)	SEP, DCI, and CCC	Students select among statements that best explain phenomena. Can use CCC in statements that students must evaluate.

Item Type	Targeted Dimensions	Description
Constructed Response	SEP (best for 3–8), DCI, CCC	Allows students to explain, using their own words, their understanding of phenomena. Can instruct students to use knowledge of particular CCC in response.
Two-part Items (EBSR)	SEP (2–8), DCI, and CCC	Effective at identifying claim and reasoning to support claim. Need to identify item type constraints (SR only or SR/TEI).
Inline Choice	SEP, DCI, and CCC	Sentence completion with drop-down menus.
Table Matching	SEP (1, 3, and 4) and DCI	Classification of images or objects.
Hot Spot/Hot Text	SEP (1–5) and DCI	Identification of specific aspects of models or diagrams that meet criteria.
Ordering	SEP, DCI, and CCC	Ordering images or text into correct sequence to describe process.

Once the available item types have been identified, scoring must also be taken into consideration, as each item type has specific scoring needs. This is especially true for TEI types, as certain item types may warrant partial credit depending on the complexity of the interactions. For example, in item types that allow students to sort responses based on specific criteria, partial credit may be warranted depending on the portion of the student response that is correct. The point values associated with each item type can also be adjusted depending on the number of interactions and complexity for that item type.

The same item type may function differently across platforms. Limitations or restrictions for each item type within a given platform should be identified early in test design to avoid development of items that are not supported due to platform limitations. Special considerations on item type limitations/restrictions should be given if item development occurs in one platform but the assessment is delivered in another. For example, a development platform may support draggers and drop bays of different sizes, but the delivery platform may only support single-sized draggers or drop bays within an item.

Establishing Clear Alignment Expectations – Item clusters must align to their intended targets in order to serve as a valid measure of what students know and can do relative to the multidimensional standards they are intended to assess. As such, alignment expectations must be made clear well in advance of initiating item cluster development.

Degrees of alignment for individual items or item clusters may range from implicit or "partial" to explicit or "strong" for any particular dimension. The minimum alignment expectations (e.g., expected degree of alignment) should be determined prior to any significant development efforts and may warrant the development of one or more prototypes prior to development to facilitate alignment expectation conversations.

Alignment expectations at multiple levels must be considered when developing item clusters. At the onset, the selected PEs and the phenomenon must ensure that appropriate alignment opportunities are afforded. Second, the context (stimulus) within which the phenomenon is presented must include multiple opportunities for assessing the intended dimensions. Third, individual items must gather appropriate evidence of students' proficiency with their intended alignments (dimensions). Finally, the alignments of all of the items within the item cluster must be considered in totality to ensure an appropriate alignment of the PE/PE Bundle and its related dimensions is achieved. For more detailed information on alignment considerations, see the companion paper "Alignment Considerations for NGSS Assessments" (CSAI, 2017a).

Training: Moving Beyond Traditional Science Assessment – Customized training prior to development work, particularly for writers and editors, is necessary to ensure high-quality item clusters by calibrating teams prior to development. Training should include a thorough review of all guiding and process documentation as well as a thorough review of available sample items (e.g., prototypes, assessment guides, practice tests). Training need not assume familiarity with three-dimensional science standards or the specific instantiation of them. Multiple trainings on various aspects of alignment (e.g., item to individual dimension, item types best suited to particular dimensions, cohesiveness of the stimulus and item set) is recommended so that a deeper understanding of alignment expectations is present at the earliest stages of the item cluster authoring process. For item cluster training in particular, initial focus should be on how to develop stimuli (i.e., contexts within which to present phenomena) to support a variety of interconnected and robust items that are multidimensional.

While initial training is important, follow-up trainings are equally as important to ensure consistency as well as hone specific aspects of development at each level across the project. Item writers and editors benefit from additional training after initial assignments are completed to focus on more complex nuances related to improving quality and alignments, as well as to incorporate feedback from editors, proofreaders, and clients.

Developing Guiding Documentation – Establishing a clear process workflow and providing the necessary guiding documentation prior to development is essential in ensuring both the quality and consistency of an assessment.

Guiding documents outlining project-specific details should be developed to the extent possible prior to beginning item development. These should include, but are not limited to, the following:

- Principles of Universal Design
- Accommodations available within the delivery platform (e.g., color theming, text-to-speech, font size)
- Development workflow (e.g., role specification, stages of the item development process: feedback incorporation, graphics development, proofing, final checks)
- Style Guide and Specifications
  - Graphics (e.g., labels, copyrights, size limitations, colors, fonts)
  - Item types (e.g., best use cases, limitations, score point ranges, partial scoring)
  - Standardized or preferred TEI directions

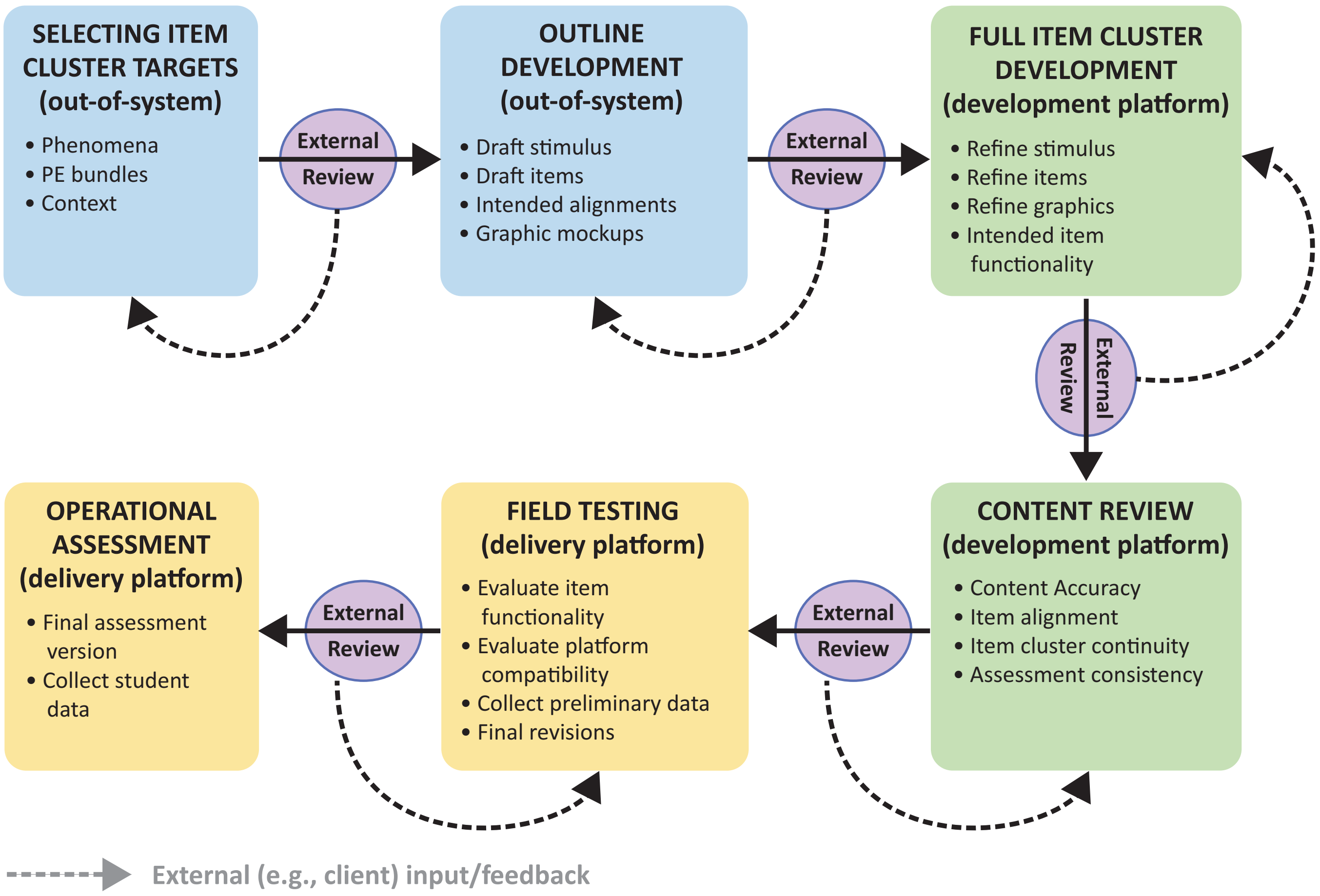
Developing these documents prior to item development helps establish clear expectations among all parties. Additionally, the documents will then serve as reference guides which can and should be updated throughout both the development and review of item materials. It is vital that these documents identify and communicate as early as possible any known limitations that may result due to platform restrictions (e.g., screen size of target administration device, scrolling behavior, color availability in development or administration platforms, defined use for dragger number and functionality, availability of images or text for specific item types) to avoid investiture into items that are not supported. Further efficiencies can then be gained by using specialized roles to complete the development work. For example, item writers and editors working together with graphic designers can be useful in generating a more refined version of an item cluster based on an approved outline to reduce inefficiencies. This allows content specialists to better focus their reviews on issues related to item alignment, content accuracy, and the overall design of the assessment.

Developing Item Clusters at Scale – Item clusters often provide a major benefit in being able to more fully assess PEs and their related dimensions (DCIs, SEPs, CCCs). This has resulted in an increased demand for the development of item clusters for use in large-scale summative assessments. The step-by-step process of developing individual item clusters is discussed in the companion paper "Design and Development of Multidimensional Science Item Clusters" (CSAI, 2017b). It is a complex process, and a number of unique challenges emerge when attempting to scale the process for the item cluster development needs of large assessment programs.

At the onset, the context of each stimulus should be robust enough to support the development of items targeting the intended alignments. Stimuli should be carefully written to provide only the information that students need to contextualize the targeted phenomena. It is vital that prerequisite content knowledge (i.e., DCIs) not be included in the stimulus as it would negatively impact measurement goals of the PE and its associated dimensions. Considerations must be given to ensure the stimulus is grade-level appropriate (determined by readability algorithms, such as Lexile or ATOS) and provides only the information relevant to the phenomenon (i.e., avoids providing information used to determine the correct answers in other items or item clusters). Both the content and language of a stimulus must then be evaluated against the individual items in the cluster to prevent cueing as well as to ensure a cohesive, focused continuity throughout the item cluster. Scaffolding can be used to provide an entry point to the item cluster as well as to vary the difficulty of individual items as the student progresses through more cognitively demanding items within the cluster.

Considering the significant time investment needed to develop item clusters in comparison to traditional standalone items, one major challenge is ensuring that the developed item clusters continue to meet design expectations for all stakeholders throughout the development process. Introducing multiple review opportunities at each major stage in the development process, as illustrated in Figure 1, is critical to ensuring that the item cluster is deemed acceptable and appropriate for a specific state assessment. In doing so, any major design adjustments that may be needed can be incorporated as early in the process as possible in order to reduce development costs, while ensuring that item clusters continue to meet client expectations. The general development process highlighting specific external review periods is outlined in Figure 1.

Figure 1: Developing Item Clusters at Scale


SELECTING ITEM CLUSTER TARGETS (out-of-system)
• Phenomena
• PE bundles
• Context
External Review
OUTLINE DEVELOPMENT (out-of-system)
• Draft stimulus
• Draft items
• Intended alignments
• Graphic mockups
External Review
FULL ITEM CLUSTER DEVELOPMENT (development platform)
• Refine stimulus
• Refine items
• Refine graphics
• Intended item functionality
External Review
CONTENT REVIEW (development platform)
• Content Accuracy
• Item alignment
• Item cluster continuity
• Assessment consistency
External Review
FIELD TESTING (delivery platform)
• Evaluate item functionality
• Evaluate platform compatibility
• Collect preliminary data
• Final revisions
External Review
OPERATIONAL ASSESSMENT (delivery platform)
• Final assessment version
• Collect student data
External (e.g., client) input/feedback

Developing outlines prior to full development, especially for item clusters, is also recommended for early evaluation purposes. Outlines should include enough information to highlight the phenomena, stimuli, and intended alignments (e.g., how targeted dimensions will be integrated into the items, rough context around which the phenomena are presented) for approval prior to the full development of an item cluster or set of standalone items. Development time can then be conserved in cases where large-scale changes, such as revising the context of an item cluster outline, are identified early on in the process. Graphic mockups for images, diagrams, graphs, and tables should also be included in draft outlines to better demonstrate how the targeted PE dimensions, particularly SEPs and CCCs, will be evaluated for any given item prior to full development.

As the number of developed item clusters grows over time, tracking the selected phenomena and stimuli contexts across a project is essential for ensuring that no single phenomenon is overrepresented in multiple item clusters. This phenomena and context tracking can then be used to help guide early item cluster development, especially in the target selection and outline development stages, in subsequent assessment development cycles.

CONCLUSION

Item clusters can be effective measurement instruments of multidimensional science learning, but only when thoughtfully designed, authored, and refined. A strongly aligned, effectively scaffolded, and deliberately structured item cluster will elicit the necessary evidence of a student's ability to apply all three dimensions of a targeted PE or PE bundle. The careful selection of a focal phenomenon and associated stimulus will ensure that the item cluster has the necessary foundational elements to assess all the required skills and knowledge. Furthermore, the time invested in outlining an item cluster will provide ample opportunity to calibrate on alignment expectations, level of student engagement, and the necessary functionality to deliver and score the intended components of an item cluster. It is through this proactive and forward-thinking process that quality multidimensional item clusters are achieved.

REFERENCES

Bybee, R. W. (2015). *The BSCS 5E Instructional Model: Creating Teachable Moments*. Arlington, Virginia: NSTA Press.

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). *Design and Development of Multidimensional Science Item Clusters.* 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 (NRC). (2012). *A framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

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.

Pellegrino, J. W. (2016). *21st Century Science Assessment: The Future Is Now.* (SRI Education White Paper). Menlo Park, CA: SRI International.

Penuel, W. R., Harris, C. J., & DeBarger, A. H. (2015). *Implementing the Next Generation Science Standards.* Phi Delta Kappan, 96(6), 45-49.

Ruiz-Primo, M. A., DiBello, L., & Solano-Flores, G. (2014). *Supporting the Implementation of the Next Generation Science Standards (NGSS) through Research: Assessment.* Retrieved from https://narst.org/ngsspapers/assessment.cfm

The Council of Chief State School Officers (CCSSO). (2015). *Science Assessment Item Collaborative (SAIC) Assessment Framework*. Washington, DC: Council of Chief State School Officers.

Whitehouse, M. (2014). *Using a Backward Design Approach to Embed Assessment in Teaching.* (SRI Education White Paper). Menlo Park, CA: SRI International.

Wiggins, G. P., McTighe, J., Kiernan, L. J., Frost, F., & Association for Supervision and Curriculum Development. (1998). *Understanding by Design.* Alexandria, Va: Association for Supervision and Curriculum Development.